ON POOLING DATA AND CHOICE OF REGRESSION PREDICTION MODELS

by

Hsien Teh Lin

Technical Report No. 246

University of Minnesota
Minneapolis, Minnesota

April, 1975

ACKNOWLEDGEMENTS

First of all, I wish to express my sincere thanks to Professor Robert J. Buehler, my adviser, for his encouragement and competent guidance through which this thesis has been written. He suggested the problem and spent innumerable hours in discussion with me. His help was invaluable.

I also want to thank Ms. Claudia Tysdal for her skilled typing.

TABLE OF CONTENTS

CHAPTER 0

INTRODUCTION

In practice we often encounter situations where a given population can be divided into several subpopulations by stratification. Suppose population π can be stratified into two subpopulations, say π_i (associated with parameter θ_i), $i = 1,2$. For example, a given human population can be divided into males and females. Suppose our major interest is in θ_1. Then our inference is usually based on the observations from π_1. Are observations from π_2 helpful to our inference about θ_1? Shall we restrict our attention only to the observations from π_1, or shall we ignore the stratification and pay attention to observations from π_2 also? The present thesis discusses this problem from several different points of view.

In Chapter 1, we discuss this problem in the case that no past data are available and our major interest is to find an estimator for θ_1 by using mean square error criterion. We consider a sample of fixed size n for estimating θ_1. We demonstrate that in both Bernoulli and Poisson populations, where mean and variance are related, sometimes we can find a better estimator for θ_1 by ignoring stratifications when we take the sample. That is, n observations from π_1 and π_2 pooled together are sometimes better than n observations from π_1. However, for the case of constant variance we don't gain by ignoring stratifications when we take the sample. The reason is that when mean

is related to variance there is a trade off between bias introduced by failing to stratify and a possibly increased variance resulting from stratifications.

The next question is what we shall do when we have past data at our disposal. In this case "ignoring stratifications" is equivalent to "pooling of data". It is well known that sometimes we can find a better estimator for θ_1 by pooling data from π_1 with data from π_2, since pooling of data will provide us with more observations. The problem of pooling data has been discussed quite extensively in the literature. Earlier work applies the method of preliminary tests of significance to the problem of pooling data. Suppose we make a preliminary test of the hypothesis that $\theta_1 = \theta_2$. If we accept the hypothesis, we pool the data; otherwise, we don't pool the data.

The idea of "ignoring stratifications" is related to the problem of choice of a regression prediction model. In standard regression models where the distribution of a dependent variable Y depends on several independent variables X_i it is well known that the mean square error of a predicted future observation may be smaller when it is based on a "deleted model" (with some of the X_i deleted) than when the full-model predictor is used.

In Chapter 2, we briefly discuss literature on preliminary tests of significance. We also briefly discuss the problem of choosing a regression prediction model. The relationship between these two problems is established in the sense that the always-pooled, the never-pooled and the sometimes-pooled estimators

correspond respectively to what we call the deleted-model, the full-model and the conditional predictors.

In Chapter 3, under the situation of availability of past data, we propose a pooling rule different from the one based on a preliminary test. In the case of two binomial populations we propose a sometimes-pooled estimator based on a linearly approximated pooling region. We carry out a numerical study to compare the performance of this estimator to that of the sometimes-pooled estimator using a preliminary test (in the Kale-Bancroft (1967) sense). The numerical results indicate that when the difference of two parameters is large the sometimes-pooled estimator based on the linearly approximated pooling region has a better performance as measured by mean square error. We also briefly discuss the properties of the always-pooled, the never-pooled and the sometimes-pooled estimators in normal and Poisson cases.

In Chapter 4, under the situation of availability of past data we introduce a new concept that puts the always-pooled, the never-pooled and the sometimes-pooled estimators into "action". We introduce the idea of treatment decision rule in connection with a pooling decision rule. We define the always-pool, the never-pool and the sometimes-pool treatment decision rules according to the corresponding estimators being used. We find that the pooling of data is irrelevant in this framework in the sense that the sometimes-pool treatment decision rule and the never-pool treatment decision rule always make the same treatment decision.

This result holds in the following cases:

two binomial populations and two treatments, arcsine square root transformation, two binomial populations and s treatments, two r-variate multinomial populations and two treatments and predictions in two normal populations.

In Chapter 5, we generalize this result in the three populations case.

In Chapter 6, we also discuss the problem of choosing a regression prediction model in a similar fashion when we discretize the problem. When we discretize the regression prediction problem, in the case of equal losses the deletion of independent variables is irrelevant in the sense that the full-model prediction decision rule and the sometimes deleted-model prediction decision rule always make the same prediction decision. In the case of unequal losses a necessary and sufficient condition for which the sometimes deleted-model prediction decision rule to be "better" than the full-model prediction decision rule is derived.

In Chapter 7, we still assume the availability of past data. We consider a standard Bayesian approach with our prior knowledge of parameters expressed in probabilistic form. In normal and regression cases we derive a necessary and sufficient condition for the Bayes risk associated with the always-pooled estimator (the deleted-model predictor) to be less than the Bayes risk associated with the never-pooled estimator (the full-model predictor). The Bayes estimators are found in normal, binomial and Poisson cases.

The Bayes predictor is also found in the regression case. Again, the relationship between the problem of pooling data and the problem of choosing a regression prediction model is established. In binomial and Poisson cases, even under the assumption of independent priors, a linear combination of the never-pooled and the always-pooled estimators is a "better" estimator than the never-pooled, although not the best.

CHAPTER I

THE PROBLEM OF ESTIMATION IN A STRATIFIED POPULATION WHEN PAST DATA ARE NOT AVAILABLE

1.1 Introduction.

In practice we often have situations that a given population can be divided into several subpopulations by stratifications. For example, a given human population can be divided into four subpopulations, say young men, young women, old men and old women, by two stratifications, namely age and sex. Suppose that subpopulation i is associated with parameter p_i. For example, the old male subpopulation is associated with a Bernoulli random variable Y_i, where $Y_i = 1$ with probability p_i if he is sick; and $Y_i = 0$ with probability $1 - p_i$ if he is healthy. Our problem is to estimate p_i when no past data are available. We take a sample of n observations to estimate p_i.

We consider different sampling procedures, which depend on whether or not we ignore stratifications. We can take n observations from the subpopulation i which is associated with p_i, the parameter of interest. Or we can ignore one stratification and take n observations randomly from all the subpopulations associated with this stratification. Or we can ignore more stratifications and take n observations randomly from the subpopulations associated with these ignored stratifications. We demonstrate that sometimes we can gain more by ignoring stratifications when we take the sample.

Among different sampling procedures we look for the best one according to a certain criterion. The two criteria that we will

consider are "mean square error" and "expected penalty". It turns out that both criteria yield the same result. Also we have three assumptions concerning sample sizes from each subpopulation, namely, (i) uniform discrete distribution on a simplex space, (ii) multinomial distribution with parameters uniformly distributed on a simplex space, (iii) multinomial distribution with equal probabilities. It turns out that (i) and (ii) give the same result. Formulas are derived according to a certain sampling assumption and a certain criterion to tell us which sampling procedure is the best. The Bernoulli, the Poisson and the constant variance cases are discussed separately.

Section 1.2 gives us some general theorems and a lemma which will be useful in our later discussions. Section 1.3 discusses the case that a given population is divided into two subpopulations by one stratification. Section 1.4 discusses the case that a given population is divided into four subpopulations by two stratifications and parameters are under "additive" restrictions. Section 1.5 discusses the same case but without "additive" restrictions.

To estimate the mean associated with a certain subpopulation, we demonstrate that in both Bernoulli and Poisson cases we can sometimes gain by ignoring one or two stratifications when we take the sample. We note that in both cases the variance is related to the mean. But in the constant variance case, where mean and variance are not related, we can never gain by ignoring either one or both stratifications. Normal distributions furnish an example.

1.2. Some useful theorems.

The following theorems and lemma will be needed in our later discussions.

Definition 2.1. We say that a t-variate $(t \geq 2)$ integer random variable $(n_1, n_2, \ldots, n_t)$ has a uniform discrete distribution on a simplex space if it satisfies the equation $\sum n_i = n$, where n is a fixed integer and each solution of the equation has the same probability.

Theorem 2.1. Suppose that a t-variate $(t \geq 2)$ integer random variable $(n_i,\ i = 1,2,\ldots,t)$ has a uniform discrete distribution on a simplex space. Then

(i) $E(n_i) = n/t$, for all i,

(ii) $\mathrm{Var}(n_i) = [(t-1)n^2 + t(t-1)n]/[t^2(t+1)]$, for all i,

(iii) $\mathrm{Cov}(n_i, n_j) = (-n^2-tn)/[t^2(t+1)]$, for all $i \neq j$.

Proof: See Appendix A.

Theorem 2.2. Let $(n_i,\ i = 1,2,\ldots,t)$ be distributed as the t-1 variate multinomial distribution with parameters $(n;\ q_i,\ i = 1,2,\ldots,t)$, i.e.,

$$P(n_1, n_2, \ldots, n_t) = \frac{n!}{n_1!\, n_2! \cdots n_t!}\ q_1^{n_1}, q_2^{n_2}, \ldots, q_t^{n_t},$$

where $n_i \geq 0$, $\sum n_i = n$, $0 < q_i < 1$, $\sum q_i = 1$. Assume that the prior distribution of $(q_i,\ i = 1,2,\ldots,t)$ is uniform on the simplex $(q_i > 0,\ \sum q_i = 1)$. Then

(i) $E(n_i)$, $Var(n_i)$ and $Cov(n_i,n_j)$ are as stated in Theorem 2.1.,

(ii) The 1st and the 2nd moments of n_i's under the above assumptions are the same as they are under assumptions in Theorem 2.1.

Proof: See Appendix A.

Theorem 2.3. Suppose Y_{ij} are independent Bernoulli random variables with parameter p_i, i.e., $Y_{ij} = 1$ with probability p_i; $Y_{ij} = 0$, with probability $1-p_i$. Let $X = \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}$, where $(n_i, i = 1,2,\ldots,t)$ is a t-variate integer random variable with a uniform discrete distribution on a simplex space (as defined in Definition 2.1). Then

(i) $EX = n(\sum p_i)/t$,

(ii) $Var\ X = n(\sum p_i)/t + \{[(t-1)n^2-2tn](\sum p_i^2)-2(n^2+tn)(\sum_{i<j} p_i p_j)\}/[t^2(t+1)]$.

Proof: See Appendix A.

Theorem 2.3'. Suppose $Y_{ij} \sim$ Poisson (λ_i). Let $X = \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}$, where $(n_i, i = 1,2,\ldots,t)$ is a t-variate integer random variable with a uniform discrete distribution on a simplex space (as defined in Definition 2.1) and Y_{ij}'s are independent. Then

(i) $EX = n(\sum \lambda_i)/t$,

(ii) $VarX = n(\sum \lambda_i)/t + [(n+t)(t-1)n(\sum \lambda_i^2) - 2(n^2+tn)(\sum_{i<j} \lambda_i \lambda_j)]/[t^2(t+1)]$.

Proof: Similar to the proof of Theorem 2.3.

Theorem 2.3''. Suppose Y_{ij} has an arbitrary distribution with mean μ_i and variance σ^2. Let $X = \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}$, where $(n_i, i = 1,2,\ldots,t)$ is a t-variate integer random variable with a uniform discrete distribution on a simplex space (as defined in Definition 2.1) and Y_{ij}'s are independent. Then

(i) $EX = n(\sum \mu_i)/t$,

(ii) $VarX = n\sigma^2 + [(n+t)(t-1)n(\sum \mu_i^2) - 2(n^2+tn)(\sum_{i<j} \mu_i \mu_j)]/[t^2(t+1)]$.

Proof: Similar to the proof of Theorem 2.3.

Theorem 2.4. Suppose Y_{ij} are independent Bernoulli random variables with parameter p_i. Let $X = \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}$, where n_i's are random variables distributed as multinomial with parameters $(n ; q_i, i = 1,2,\ldots,t)$. Then

(i) $EX = n \sum q_i p_i$,

(ii) $VarX = n \sum q_i p_i (1 - q_i p_i) - 2n \sum_{i<j} q_i q_j p_i p_j$.

Proof: Similar to the proof of Theorem 2.3.

Corollary 2.1. In addition to the assumptions in Theorem 2.4, we assume that the prior distribution of $(q_i, i = 1,2,\ldots,t)$ is uniform on the simplex $(q_i > 0, \sum q_i = 1)$. Then EX and VarX are the same as stated in Theorem 2.3.

Proof: See Appendix A.

Remark: Let $X = \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}$, where Y_{ij}'s are independent Bernoulli random variables and n_i's are random variables. The 1st and the 2nd moments of X (or the mean and the variance of X) under the assumption that $(n_i, i = 1,2,\ldots,t)$ has a uniform discrete distribution

on a simplex space are the same as those of X under the assumption that n_i's are multinomial with parameters uniformly distributed on the simplex space. Actually, this striking fact is a consequence of Theorem 2.2. This fact makes it possible to conclude in later discussions that two different sampling assumptions will yield the same result.

Corollary 2.2. In addition to the assumptions in Theorem 2.4, we assume that $q_i = 1/t$, for all i. Then

(i) $EX = n(\sum p_i)/t$,

(ii) $VarX = n(\sum p_i)/t - [n(\sum p_i^2) + 2n(\sum_{i<j} p_i p_j)]/t^2$.

Proof: It follows directly from Theorem 2.4.

Theorem 2.4': Suppose $Y_{ij} \sim$ Poisson (λ_i). Let $X = \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}$, where n_i's are random variables distributed as multinomial with parameters $(n\ ;\ q_i,\ i = 1,2,\ldots,t)$ and Y_{ij}'s are independent. Then

(i) $EX = n \sum q_i \lambda_i$,

(ii) $VarX = n[\sum q_i \lambda_i + \sum \lambda_i^2 q_i(1-q_i) - 2 \sum_{i<j} \lambda_i \lambda_j\ q_i q_j]$.

Proof: Similar to the proof of Theorem 2.3.

Corollary 2.1': In addition to the assumptions in Theorem 2.4', we assume that the prior distribution of $(q_i,\ i = 1,2,\ldots,t)$ is uniform on the simplex $(q_i > 0,\ \sum q_i = 1)$. Then EX and VarX are the same as stated in Theorem 2.3'.

Proof: Similar to the proof of Corollary 2.1.

Corollary 2.2'. In addition to the assumptions in Theorem 2.4', we assume that $q_i = 1/t$, for all i. Then

(i) $EX = n(\sum \lambda_i)/t$,

(ii) $VarX = n(\sum \lambda_i)/t + [n(t-1)(\sum \lambda_i^2) - 2n(\sum_{i<j} \lambda_i \lambda_j)]/t^2$.

Proof: It follows directly from Theorem 2.4'.

Theorem 2.4''. Suppose Y_{ij} has an arbitrary distribution with mean μ_i and variance σ^2. Let $X = \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}$, where n_i's are random variables distributed as multinomial with parameters $(n; q_i, i = 1,2,\ldots t)$ and Y_{ij}'s are independent. Then

(i) $EX = n \sum q_i \mu_i$,

(ii) $VarX = n[\sigma^2 + \sum \mu_i^2 q_i(1-q_i) - 2 \sum_{i<j} \mu_i \mu_j q_i q_j]$.

Proof: Similar to the proof of Theorem 2.3.

Corollary 2.1''. In addition to the assumptions in Theorem 2.4'', we assume that the prior distribution of $(q_i, i = 1,2,\ldots,t)$ is uniform on the simplex $(q_i > 0, \sum q_i = 1)$. Then EX and VarX are the same as stated in Theorem 2.3''.

Proof: Similar to the proof of Corollary 2.1.

Corollary 2.2''. In addition to the assumptions in Theorem 2.4'', we assume that $q_i = 1/t$, for all i. Then

(i) $EX = n(\sum \mu_i)/t$,

(ii) $VarX = n\sigma^2 + [n(t-1)(\sum \mu_i^2) - 2n(\sum_{i<j} \mu_i \mu_j)]/t^2$.

Proof: It follows directly from Theorem 2.4''.

Lemma 2.1. Let $\hat{p}_i$, $i = 1,2$ be the estimators of p. Then

$$E(\hat{p}_1-p)^2 - E(\hat{p}_2-p)^2 = (E\hat{p}_1^2 - E\hat{p}_2^2) - 2p(E\hat{p}_1 - E\hat{p}_2) .$$

Remark: Lemma 2.1 will make us able to conclude in later discussions that in the Bernoulli case two criteria, namely "mean square error" and "expected penalty", always yield the same result.

1.3. 1 x 2 case.

1.3.1. Bernoulli case.

Suppose a given population can be divided into two subpopulations, say A and $\bar{A}$, by a certain stratification. Suppose each subpopulation is associated with a Bernoulli random variable with a parameter as indicated respectively in Table 3.1 and Table 3.2.

A	$\bar{A}$
Y_{1j}	Y_{2j}

Table 3.1

A	$\bar{A}$
p	$p+\tau$

Table 3.2

We define

$$\begin{cases} Y_{1j} = 1, & \text{with probability } p \\ \quad = 0, & \text{with probability } 1-p \end{cases}$$

and

$$\begin{cases} Y_{2j} = 1, & \text{with probability } p+\tau \\ \quad = 0, & \text{with probability } 1-p-\tau . \end{cases}$$

We want to estimate p when no past data are available. We have to take a sample of n observations to estimate p. We have the

following two different sampling procedures, depending on if we ignore the stratification or not.

Sampling procedure 1. Consider stratification. We take n observations $(Y_{1j}, j = 1,2,\ldots,n)$ randomly from the subpopulation A. Let

$$\hat{p}_1 = \sum_{j=1}^{n} Y_{1j}/n$$

be an estimator of p.

Sampling procedure 2. Ignore stratification. We take n observations randomly from both subpopulations A and $\bar{A}$. Let the n observations be denoted by Y_{ij}, $j = 1,2,\ldots,n_i$, $i = 1,2$, $n_1 + n_2 = n$. Let

$$\hat{p}_2 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}/n$$

be an estimator of p.

We want to know when $\hat{p}_2$ is better than $\hat{p}_1$, and vice versa. The two criteria that we will consider are as follows:

Criterion 1. Mean square error (M.S.E.). Let $\hat{p}_i$ be an estimator of p. Then

$$M.S.E.(\hat{p}_i) = E(\hat{p}_i - p)^2 = \text{Var } \hat{p}_i + (E\hat{p}_i - p)^2 .$$

Criterion 2. Expected penalty (E.P.). Let $\hat{p}_i$ be an estimator of p. Then we define that

$$\text{penalty} = \begin{cases} \hat{p}_i^2 & , \text{ if } Y_{1j} = 0 \\ (1-\hat{p}_i)^2, & \text{if } Y_{1j} = 1. \end{cases}$$

Then

$$E.P.(\hat{p}_i) = E(\hat{p}_i)^2(1-p) + E(1-\hat{p}_i)^2 p = E\hat{p}_i^2 - 2pE\hat{p}_i + p.$$

By Lemma 2.1, it is easy to see that

$$M.S.E.(\hat{p}_2) - M.S.E.(\hat{p}_1) = E.P.(\hat{p}_2) - E.P.(\hat{p}_1) .$$

If we take n observations randomly from t subpopulations (in this particular case, $t = 2$) in such a manner that n_i is from the ith subpopulation and $\sum_{i=1}^{t} n_i = n$, it is clear that n_i is a random variable. We have the following assumptions concerning the distribution of n_i.

Assumption 1. Uniform discrete distribution on a simplex space (as defined in Definition 2.1).

Assumption 2. Multinomial distribution with parameters uniformly distributed on a simplex space.

Assumption 3. Multinomial distributions with equal probabilities.

Because of Corollary 2.1, we will get the same $M.S.E.(\hat{p}_i)$, $i = 1,2$, either under Assumption 1 or 2.

Now under Assumption 1 or 2, we apply Corollary 2.1 to compute $M.S.E.(\hat{p}_i)$; and we get

(i) $M.S.E.(\hat{p}_1) = p(1-p)/n$,

(ii) $M.S.E.(\hat{p}_2) = p(1-p)/n + \tau(2n\tau - 2\tau - 6p+3)/(6n)$.

Therefore, under Criterion 1 or 2 and Assumption 1 or 2, $\hat{p}_2$ is better than $\hat{p}_1$ (in the sense that $M.S.E.(\hat{p}_2) < M.S.E.(\hat{p}_1)$) if and only if either (i) $0 < \tau < 3(p - \frac{1}{2})/(n-1)$ and $p > \frac{1}{2}$ or (ii) $3(p-\frac{1}{2})/(n-1) < \tau < 0$ and $p < \frac{1}{2}$.

Next, under Assumption 3, we can apply Corollary 2.2 to compute M.S.E.$(\hat{p}_i)$; and we get

(i) M.S.E.$(\hat{p}_1) = p(1-p)/n$,

(ii) M.S.E.$(\hat{p}_2) = p(1-p)/n + \tau[n\tau-\tau-4p+2]/(4n)$.

Consequently, under Criterion 1 or 2 and Assumption 3, $\hat{p}_2$ is better than $\hat{p}_1$ (in the sense that M.S.E.$(\hat{p}_2)$ < M.S.E.$(\hat{p}_1)$) if and only if either (i) $0 < \tau < 4(p-\frac{1}{2})/(n-1)$ and $p > \frac{1}{2}$ or (ii) $4(p-\frac{1}{2})/(n-1) < \tau < 0$ and $p < \frac{1}{2}$.

1.3.2. Poisson case.

Next, we consider the case that each subpopulation is associated with a Poisson random variable with a parameter as indicated respectively in Talbe 3.3 and Table 3.4.

A	$\bar{A}$
Y_{1j}	Y_{2j}

Table 3.3

A	$\bar{A}$
λ	$\lambda+\tau$

Table 3.4

That is, $Y_{1j} \sim$ Poisson (λ) and $Y_{2j} \sim$ Poisson $(\lambda+\tau)$.

Our problem is to estimate λ when no past data are available. We have to take a sample of n observations to estimate λ. Consequently, we have two estimators for λ, namely,

$\hat{\lambda}_1 = \sum_{j=1}^{n} Y_{1j}/n$ (under Sampling procedure 1)

$\hat{\lambda}_2 = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} Y_{ij}/n$ (under Sampling procedure 2).

Again, we want to know when $\hat{\lambda}_2$ is better than $\hat{\lambda}_1$ according to a

certain criterion. In the Poisson case it is meaningful to consider only Criterion 1.

Under Assumption 1 or 2 we can apply Corollary 2.1' to compute $M.S.E.(\hat{\lambda}_i)$; and we get

(i) $M.S.E.(\hat{\lambda}_1) = \lambda/n$,

(ii) $M.S.E.(\hat{\lambda}_2) = \lambda/n + \tau(2n\tau + \tau + 3)/(6n)$.

It follows that under Assumption 1 or 2, $M.S.E.(\hat{\lambda}_2) < M.S.E.(\hat{\lambda}_1)$ if and only if $-3/(2n + 1) < \tau < 0$.

Similarly, under Assumption 3, we can apply Corollary 2.2' to compute $M.S.E.(\hat{\lambda}_i)$; and we get

(i) $M.S.E.(\hat{\lambda}_1) = \lambda/n$,

(ii) $M.S.E.(\hat{\lambda}_2) = \lambda/n + \tau(n\tau + \tau + 2)/(4n)$.

It follows that under Assumption 3, $M.S.E.(\hat{\lambda}_2) < M.S.E.(\hat{\lambda}_1)$ if and only if $-2/(n+1) < \tau < 0$.

We note that under either assumption, the necessary and sufficient condition for $M.S.E.(\hat{\lambda}_2) < M.S.E.(\hat{\lambda}_1)$ does not involve λ.

1.3.3. Constant variance case.

Finally, we consider the case that each subpopulation is associated with a random variable Y_{ij} (indicated in Table 3.5) having an arbitrary distribution. Its mean is indicated in Table 3.6.

A	$\bar{A}$
Y_{1j}	Y_{2j}

Table 3.5

A	$\bar{A}$
μ	$\mu+\tau$

Table 3.6

We assume that $\mathrm{Var}(Y_{ij}) = \sigma^2$, for all i and j.

Again, our problem is to estimate μ when no past data are available. We have to take a sample of n observations to estimate μ. We have two estimators for μ, namely,

$\hat{\mu}_1 = \sum_{j=1}^{n} Y_{1j}/n$ (under Sampling procedure 1)

$\hat{\mu}_2 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}/n$ (under Sampling procedure 2).

We want to know when $\hat{\mu}_2$ is better than $\hat{\mu}_1$ according to Criterion 1. In this case, it is meaningful to consider only Criterion 1.

Under Assumption 1 or 2, we can apply Corollary 2.1'' to get

(i) $\mathrm{M.S.E.}(\hat{\mu}_1) = \sigma^2/n$,

(ii) $\mathrm{M.S.E.}(\hat{\mu}_2) = [6\sigma^2 + (2n+1)\tau^2]/(6n)$.

Under Assumption 3, we can apply Corollary 2.2'' to get

(i) $\mathrm{M.S.E.}(\hat{\mu}_1) = \sigma^2/n$,

(ii) $\mathrm{M.S.E.}(\hat{\mu}_2) = [4\sigma^2 + (n+1)\tau^2]/(4n)$.

Under either assumption, $\mathrm{M.S.E.}(\hat{\mu}_1) \leq \mathrm{M.S.E.}(\hat{\mu}_2)$. In other words, in the constant variance case, we don't gain by ignoring stratification. Normal distributions furnish an example.

1.4. 2 x 2 additive case.

1.4.1. Bernoulli case.

Suppose a given population can be divided into four subpopulations, say AB, $A\bar{B}$, $\bar{A}B$ and $\bar{A}\bar{B}$, by two stratifications. Suppose each subpopulation is associated with a Bernoulli random

variable with parameter p_{ij} as indicated in Table 4.1 and Table 4.2.

	A	$\bar{A}$
B	Y_{1j}	Y_{3j}
$\bar{B}$	Y_{2j}	Y_{4j}

Table 4.1

	A	$\bar{A}$
B	p_{11}	p_{21}
$\bar{B}$	p_{12}	p_{22}

Table 4.2

We define that

$$\begin{cases} Y_{1j} = 1, & \text{with probability } p_{11} \\ \quad = 0, & \text{with probability } 1-p_{11}. \end{cases}$$

Similarly defined for other Y_{ij}'s. We assume that p_{ij}'s are under "additive" restrictions, namely (i) $p_{21}-p_{11} = p_{22}-p_{12} = \tau$, (ii) $p_{12}-p_{11} = p_{22}-p_{21} = \beta$. We can rewrite Table 4.2 as follows:

	A	$\bar{A}$
B	p	p+τ
$\bar{B}$	p+β	p+β+τ

Table 4.3

We want to estimate p when no past data are available. We have to take a sample of n observations to estimate p. We have the following three sampling procedures, which depend on the stratifications we ignore.

Sampling procedure 1. Consider both stratifications. We take n observations $(Y_{1j}, j = 1,2,\ldots,n)$ randomly the subpopulation AB. Let

$$\hat{p}_1 = \sum_{j=1}^{n} Y_{1j}/n$$

be an estimator of p.

Sampling procedure 2. Ignore one stratification. Without loss of generality, assume that we ignore B stratification. We take n observations randomly from subpopulations AB and $A\bar{B}$. Let the n observations be denoted by Y_{ij}, $j = 1,2,\ldots, n_i$, $i = 1,2$, $n_1+n_2 = n$.

Let
$$\hat{p}_2 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}/n$$
be an estimator of p.

Sampling procedure 3. Ignore both stratifications. We take n observations randomly from all four subpopulations. Let the n observations be denoted by Y_{ij}, $j = 1,2,\ldots,n_i$, $i = 1,2,3,4$, $\sum n_i = n$.

Let
$$\hat{p}_3 = \sum_{i=1}^{4} \sum_{j=1}^{n_i} Y_{ij}/n$$
be an estimator of p.

We have three estimators, $\hat{p}_i$, $i = 1,2,3$, for p, depending on which sampling procedure we use. We want to know which estimator is the best. The two criteria that we will consider are Criterion 1 and 2 as indicated in Section 1.3.1. Also we have the same three assumptions concerning the distribution of n_i as indicated in Section 1.3.1.

Under Assumption 1 or 2, we apply Corollary 2.1 to compute $M.S.E.(\hat{p}_i)$; and we get

(i) $\varphi_1 = M.S.E.(\hat{p}_1) = p(1-p)/n,$

(ii) $\varphi_2 = M.S.E.(\hat{p}_2) = p(1-p)/n + \beta[2n\beta-2\beta-6p+3]/(6n),$

(iii) $\varphi_3 = M.S.E.(\hat{p}_3) = p(1-p)/n+[(3n-3)(\beta^2+\tau^2)+(\beta+\tau)(5-10p)$
$+(5-5n)\beta\tau]/(10n)$.

It follows that

$$\text{(i)}\ \varphi_1-\varphi_2 = -\beta[2n\beta-2\beta-6p+3]/(6n),$$

$$\text{(4.1)}\quad \text{(ii)}\ \varphi_1-\varphi_3 = -[(3n-3)(\beta^2+\tau^2)+(\beta+\tau)(5-10p)+(5-5n)\beta\tau]/(10n),$$

$$\text{(iii)}\ \varphi_2-\varphi_3 = \{(n-1)\beta^2-\tau[(9n-9)\tau-(15-15n)\beta+30p-15]\}/(30n).$$

Consequently, under Criterion 1 or 2 and Assumption 1 or 2, (4.1) can give us respectively a necessary and sufficient condition for which $\hat{p}_i$ is best, $i = 1,2,3$.

Under Criterion 1 or 2 and Assumption 1 or 2, we have the following interesting examples and cases.

(1) Example 1. If $n = 30$, $p = 3/4$, $\beta = 1/1000$, $\tau = 1/10$, then $\hat{p}_2$ is best.

(2) Example 2. If $n = 30$, $p = 3/4$, $\beta = 0$, $\tau = 1/1000$, then $\hat{p}_3$ is best.

(3) If $\tau = \beta = 0$, then $\text{M.S.E.}(\hat{p}_1) = \text{M.S.E.}(\hat{p}_2) = \text{M.S.E.}(\hat{p}_3)$ as we expect.

(4) As $n \to \infty$, we note that (i) $\varphi_1 \to 0$ (ii) $\varphi_2 \to \beta^2/3$. (iii) $\varphi_3 \to (3/10)(\beta^2+\tau^2)+\beta\tau/2$.

It is easy to see that $(3/10)(\beta^2+\tau^2)+\beta\tau/2 \geq (\frac{1}{2}\beta+\frac{1}{2}\tau)^2 \geq 0$. Hence $\hat{p}_1$ is a consistent estimator for p_1. Also $\hat{p}_1$ is best when n is sufficiently large. In other words, we don't gain by ignoring stratifications when the sample size is sufficiently large.

Under Assumption 3, we apply Corollary 2.2 to compute $\text{M.S.E.}(\hat{p}_i)$; and we get

$$\text{(i)}\quad \varphi_1' = \text{M.S.E.}(\hat{p}_1) = p(1-p)/n,$$

$$\text{(ii)}\quad \varphi_2' = \text{M.S.E.}(\hat{p}_2) = p(1-p)/n + \beta[n\beta-\beta-4p+2]/(4n),$$

$$\text{(iii)}\quad \varphi_3' = \text{M.S.E.}(\hat{p}_3) = p(1-p)/n + [(n-1)(\beta^2+\tau^2)+(\beta+\tau)(2-4p) + (2-2n)\beta\tau]/(4n).$$

It follows that

$$\text{(i)}\quad \varphi_1'-\varphi_2' = -\beta[n\beta-\beta-4p+2]/(4n),$$

$$\text{(4.2) (ii)}\quad \varphi_1'-\varphi_3' = -[(n-1)(\beta^2+\tau^2)+(\beta+\tau)(2-4p)+(2-2n)\beta\tau]/(4n),$$

$$\text{(iii)}\quad \varphi_2'-\varphi_3' = \tau[(1-n)(\tau-2\beta)+4p-2]/(4n).$$

Consequently, under Criterion 1 or 2 and Assumption 3, (4.2) can give us respectively a necessary and sufficient condition for which $\hat{p}_i$ is best, $i = 1,2,3$.

Under Criterion 1 or 2 and Assumption 3, we have the following interesting examples and cases.

(1) Example 3. If $n = 30$, $p = 3/4$, $\beta = 1/1000$, $\tau = 1/10$, then $\hat{p}_2$ is best.

(2) Example 4. If $n = 30$, $p = 3/4$, $\beta = 0$, $\tau = 1/1000$, then $\hat{p}_3$ is best.

(3) If $\tau = \beta = 0$, then $\text{M.S.E.}(\hat{p}_1) = \text{M.S.E.}(\hat{p}_2) = \text{M.S.E.}(\hat{p}_3)$ as we expect.

(4) $\hat{p}_1$ is a consistent estimator for p_1. Also $\hat{p}_1$ is best when n is sufficiently large.

1.4.2. Poisson case.

Next, we consider the case that each subpopulation is associated with a Poisson random variable Y_{ij} with its parameter (under "additive" restrictions) as indicated by Table 4.4 and Table 4.5.

	A	$\bar{A}$
B	Y_{1j}	Y_{3j}
$\bar{B}$	Y_{2j}	Y_{4j}

Table 4.4

	A	$\bar{A}$
B	λ	$\lambda+\tau$
$\bar{B}$	$\lambda+\beta$	$\lambda+\beta+\tau$

Table 4.5

That is, $Y_{1j} \sim$ Poisson (λ). Similarly defined for other Y_{kj}'s.

Our problem is to estimate λ when no past data are available. We have to take a sample of n observations to estimate λ. We have three estimators for λ as follows:

$$\hat{\lambda}_1 = \sum_{j=1}^{n} Y_{1j}/n \quad \text{(under Sampling procedure 1)}$$

$$\hat{\lambda}_2 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}/n \quad \text{(under Sampling procedure 2)}$$

$$\hat{\lambda}_3 = \sum_{i=1}^{4} \sum_{j=1}^{n_i} Y_{ij}/n \quad \text{(under Sampling procedure 3).}$$

We want to know which estimator is best according to Criterion 1 (indicated in Section 1.3.1).

Under Assumption 1 or 2, we apply Corollary 2.1' to compute M.S.E.$(\hat{\lambda}_i)$; and we get

$$\text{(i)} \quad \varphi_1 = \text{M.S.E.}(\hat{\lambda}_1) = \lambda/n,$$

$$\text{(ii)} \quad \varphi_2 = \text{M.S.E.}(\hat{\lambda}_2) = \lambda/n + \beta(2n\beta+\beta+3)/(6n),$$

$$\text{(iii)} \quad \varphi_3 = \text{M.S.E.}(\hat{\lambda}_3) = \lambda/n + [(3n+2)(\beta^2+\tau^2)+5n\beta\tau+5(\beta+\tau)]/(10n).$$

It follows that

$$\text{(i)}\quad \varphi_1-\varphi_2 = -\beta(2n\beta+\beta+3)/(6n),$$

$$\text{(4.3)}\quad \text{(ii)}\quad \varphi_1-\varphi_3 = -[(3n+2)(\beta^2+\tau^2)+5n\beta\tau+5(\beta+\tau)]/(10n),$$

$$\text{(iii)}\quad \varphi_2-\varphi_3 = [(n-1)\beta^2-(9n+6)\tau^2-15\tau(n\beta+1)]/(30n).$$

We note that the three equations in (4.3) do not involve λ. Consequently, under Criterion 1 and Assumption 1 or 2, (4.3) can give us respectively a necessary and sufficient condition for which $\hat{\lambda}_i$ is best, $i = 1,2,3$.

Under Criterion 1 and Assumption 1 or 2, we have the following interesting examples and cases.

(1) Example 1. If $n = 5$, $\beta = -1/5$, $\tau = 10$, then $\hat{\lambda}_2$ is best.

(2) Example 2. If $n = 5$, $\beta = -1/5$, $\tau = 0$, then $\hat{\lambda}_3$ is best.

(3) If $\tau = \beta = 0$, then $\text{M.S.E.}(\hat{\lambda}_1) = \text{M.S.E.}(\hat{\lambda}_2) = \text{M.S.E.}(\hat{\lambda}_3)$ as we expect.

(4) As $n \to \infty$, we note that (i) $\varphi_1 \to 0$ (ii) $\varphi_2 \to \beta^2/3$ (iii) $\varphi_3 \to (3/10)(\beta^2+\tau^2)+\beta\tau/2$.

It is easy to see that $(3/10)(\beta^2+\tau^2)+\beta\tau/2 \geq (\frac{1}{2}\beta + \frac{1}{2}\tau)^2 \geq 0$. Hence $\hat{\lambda}_1$ is a consistent estimator for λ_1. Also $\hat{\lambda}_1$ is best when n is sufficiently large. In other words, we don't gain by ignoring stratifications when the sample size is sufficiently large.

Under Assumption 3, we apply Corollary 2.2' to compute $\text{M.S.E.}(\hat{\lambda}_i)$; and we get

$$\text{(i)}\quad \varphi_1' = \text{M.S.E.}(\hat{\lambda}_1) = \lambda/n,$$

$$\text{(ii)}\quad \varphi_2' = \text{M.S.E.}(\hat{\lambda}_2) = \lambda/n + \beta(n\beta+\beta+2)/(4n),$$

$$\text{(iii)}\quad \varphi_3' = \text{M.S.E.}(\hat{\lambda}_3) = \lambda/n + [(n+1)(\beta^2+\tau^2)+2n\beta\tau+2(\beta+\tau)]/(4n).$$

It follows that

$$\text{(i)}\quad \varphi_1'-\varphi_2' = -\beta(n\beta+\beta+2)/(4n),$$

$$\text{(4.4)}\quad \text{(ii)}\quad \varphi_1'-\varphi_3' = -[(n+1)(\beta^2+\tau^2)+2n\beta\tau+2(\beta+\tau)]/(4n),$$

$$\text{(iii)}\quad \varphi_2'-\varphi_3' = -\tau[(n+1)\tau+2n\beta+2]/(4n).$$

Again, we note that the three equations in (4.4) do not involve λ. Under Criterion 1 and Assumption 3, (4.4) can give us respectively a necessary and sufficient condition for which $\hat{\lambda}_i$ is best, $i = 1,2,3$.

Under Criterion 1 and Assumption 3, we have the following interesting examples and cases.

(1) Example 3. If $n = 5$, $\beta = -1/5$, $\tau = 10$, then $\hat{\lambda}_2$ is best.

(2) Example 4. If $n = 5$, $\beta = -1/5$, $\tau = 0$, then $\hat{\lambda}_2$ and $\hat{\lambda}_3$ are equally good and better than $\hat{\lambda}_1$.

(3) If $\tau = \beta = 0$, then $\text{M.S.E.}(\hat{\lambda}_1) = \text{M.S.E.}(\hat{\lambda}_2) = \text{M.S.E.}(\hat{\lambda}_3)$ as we expect.

(4) $\hat{\lambda}_1$ is a consistent estimator for λ_1. Also $\hat{\lambda}_1$ is best when n is sufficiently large.

1.4.3. Constant variance case.

Finally, we consider the case that each subpopulation is associated with a random variable Y_{ij} (indicated in Table 4.6) having an arbitrary distribution. Its mean (under "additive" restrictions) is indicated in Table 4.7.

	A	$\bar{A}$
B	Y_{1j}	Y_{3j}
$\bar{B}$	Y_{2j}	Y_{4j}

Table 4.6

	A	$\bar{A}$
B	μ	$\mu+\tau$
$\bar{B}$	$\mu+\beta$	$\mu+\beta+\tau$

Table 4.7

We assume that $\text{Var}(Y_{ij}) = \sigma^2$, for all i and j.

Again, our problem is to estimate μ when no past data are available. We have to take a sample of n observations to estimate μ. Consequently, we have three estimators for μ as follows:

$\hat{\mu}_1 = \sum_{j=1}^{n} Y_{1j}/n$ (under Sampling procedure 1)

$\hat{\mu}_2 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}/n$ (under Sampling procedure 2)

$\hat{\mu}_3 = \sum_{i=1}^{4} \sum_{j=1}^{n_i} Y_{ij}/n$ (under Sampling procedure 3).

We want to know which estimator is best according to Criterion 1 (indicated in Section 1.3.1).

Under Assumption 1 or 2, we apply Corollary 2.1'' to get

(i) $\text{M.S.E.}(\hat{\mu}_1) = \sigma^2/n$,

(ii) $\text{M.S.E.}(\hat{\mu}_2) = [6\sigma^2 + (2n+1)\beta^2]/(6n)$,

(iii) $\text{M.S.E.}(\hat{\mu}_3) = [20\sigma^2 + (n+4)(\beta^2+\tau^2)]/(20n) + (\beta/2 + \tau/2)^2$.

Under Assumption 3, we apply Corollary 2.2'' to get

(i) $\text{M.S.E.}(\hat{\mu}_1) = \sigma^2/n$,

(ii) $\text{M.S.E.}(\hat{\mu}_2) = [4\sigma^2 + (n+1)\beta^2]/(4n)$,

(iii) $\text{M.S.E.}(\hat{\mu}_3) = [4\sigma^2 + (\beta^2+\tau^2)]/(4n) + (\beta/2 + \tau/2)^2$.

Under either assumption, $\text{M.S.E.}(\hat{\mu}_1) \leq \min[\text{M.S.E.}(\hat{\mu}_2), \text{M.S.E.}(\hat{\mu}_3)]$.

In the constant variance case, where mean and variance are not related, we don't gain by ignoring stratifications. Normal distributions furnish an example. However, in both Bernoulli and Poisson cases, where mean and variance are realted, sometimes we can find a better estimator for mean by ignoring one or both stratifications when we take the sample.

1.5. 2 x 2 non-additive case.

The formulation of problem and assumptions are exactly the same as in Section 1.4 except that parameters are no longer under "additive" restrictions. The relaxation of "additive" assumption will result in one more parameter entering into all the algebraic expressions already shown in Section 1.4.

In the Bernoulli case we have parameters as indicated in Table 4.2. Let $p = p_{11}$, $\beta = p_{12}-p_{11}$, $\tau = p_{21}-p_{11}$ and $\gamma = (p_{22}-p_{12})-(p_{21}-p_{11})$. Without "additive" restrictions the parameters can be rewritten as follows:

	A	$\bar{A}$
B	p	p+τ
$\bar{B}$	p+β	p+β+τ+γ

Table 5.1

Under Assumption 1 or 2, we can apply Corollary 2.1 to compute M.S.E.$(\hat{p}_i)$; and we get

$$
\begin{aligned}
&\text{(i)} \quad \text{M.S.E.}(\hat{p}_1) - \text{M.S.E.}(\hat{p}_2) = \varphi_1 - \varphi_2, \\
(5.1)\quad &\text{(ii)} \quad \text{M.S.E.}(\hat{p}_1) - \text{M.S.E.}(\hat{p}_3) = \varphi_1 - \varphi_3 - \varphi(\gamma), \\
&\text{(iii)} \quad \text{M.S.E.}(\hat{p}_2) - \text{M.S.E.}(\hat{p}_3) = \varphi_2 - \varphi_3 - \varphi(\gamma),
\end{aligned}
$$

where $\varphi_i - \varphi_j$ is the same as indicated in (4.1) and $\varphi(\gamma) = \gamma[(2n-2)\gamma - 10p + (6n-6)(\beta+\tau) + 5]/(20n)$.

Under Assumption 3, we can apply Corollary 2.2 to compute $\text{M.S.E.}(\hat{p}_i)$; and we get

$$
\begin{aligned}
&\text{(i)} \quad \text{M.S.E.}(\hat{p}_1) - \text{M.S.E.}(\hat{p}_2) = \varphi_1' - \varphi_2', \\
(5.2)\quad &\text{(ii)} \quad \text{M.S.E.}(\hat{p}_1) - \text{M.S.E.}(\hat{p}_3) = \varphi_1' - \varphi_3' - \varphi'(\gamma), \\
&\text{(iii)} \quad \text{M.S.E.}(\hat{p}_2) - \text{M.S.E.}(\hat{p}_3) = \varphi_2' - \varphi_3' - \varphi'(\gamma),
\end{aligned}
$$

where $\varphi_i' - \varphi_j'$ is the same as indicated in (4.2) and $\varphi'(\gamma) = \gamma[(n-1)\gamma - 8p + (4n-4)(\beta+\tau) + 4]/(16n)$.

Both (5.1) and (5.2) can give us a necessary and sufficient condition for which $\hat{p}_i$ is best, $i = 1,2,3$. It is easy to see that (5.1) and (5.2) are respectively identical to (4.1) and (4.2) when $\gamma = 0$.

In the Poisson case, without "additive" restrictions, we can re-write the parameters in Table 4.5 as follows:

	A	$\bar{A}$
B	λ	$\lambda+\tau$
$\bar{B}$	$\lambda+\beta$	$\lambda+\beta+\tau+\gamma$

Table 5.2

Under Assumption 1 or 2, we can apply Corollary 2.1' to compute $\text{M.S.E.}(\hat{\lambda}_i)$; and we get

$$\text{(i)}\quad \text{M.S.E.}(\hat{\lambda}_1) - \text{M.S.E.}(\hat{\lambda}_2) = \varphi_1 - \varphi_2,$$

$$\text{(5.3)}\quad \text{(ii)}\quad \text{M.S.E.}(\hat{\lambda}_1) - \text{M.S.E.}(\hat{\lambda}_3) = \varphi_1 - \varphi_3 - \varphi(\gamma),$$

$$\text{(iii)}\quad \text{M.S.E.}(\hat{\lambda}_2) - \text{M.S.E.}(\hat{\lambda}_3) = \varphi_2 - \varphi_3 - \varphi(\gamma),$$

where $\varphi_i - \varphi_j$ is the same as indicated in (4.3) and $\varphi(\gamma) = \gamma[(2n+3)\gamma + (6n+4)(\beta+\tau) + 5]/(20n)$.

Under Assumption 3, we can apply Corollary 2.2' to compute $\text{M.S.E.}(\hat{\lambda}_i)$; and we get

$$\text{(i)}\quad \text{M.S.E.}(\hat{\lambda}_1) - \text{M.S.E.}(\hat{\lambda}_2) = \varphi_1' - \varphi_2',$$

$$\text{(5.4)}\quad \text{(ii)}\quad \text{M.S.E.}(\hat{\lambda}_1) - \text{M.S.E.}(\hat{\lambda}_3) = \varphi_1' - \varphi_3' - \varphi'(\gamma),$$

$$\text{(iii)}\quad \text{M.S.E.}(\hat{\lambda}_2) - \text{M.S.E.}(\hat{\lambda}_3) = \varphi_2' - \varphi_3' - \varphi'(\gamma),$$

where $\varphi_i' - \varphi_j'$ is the same as indicated in (4.4) and $\varphi'(\gamma) = \gamma[(n+3)\gamma + (4n+4)(\beta+\tau) + 4]/(16n)$.

Both (5.3) and (5.4) can give us a necessary and sufficient condition for which $\hat{\lambda}_i$ is best, $i = 1,2,3$. It is easy to see that (5.3) and (5.4) are respectively identical to (4.3) and (4.4) when $\gamma = 0$. Also we note that equations in both (5.3) and (5.4) do not involve λ, the parameter of interest.

CHAPTER 2

THE PROBLEM OF POOLING DATA AND THE PROBLEM OF CHOOSING A REGRESSION PREDICTION MODEL

2.1. Introduction.

We have demonstrated in Chapter 1 that, under the situation of no availability of past data, to estimate a certain parameter associated with a certain subpopulation we can sometimes gain by ignoring stratifications when we take the sample. The next question is what we shall do when we have past data at our disposal. Under the situation of availability of past data, "ignoring stratifications" is equivalent to "pooling of data". The problems concerned with pooling of data have been discussed quite extensively in the literature. Also the idea of "ignoring stratifications" can be related to "choice of model" in regression theory.

This chapter gives a brief introduction to "pooling of data" and "choice of model" and their relationship.

In Section 2.2, we briefly survey literature in statistical inference using preliminary tests of significance, with special application to the problem of pooling data. In Section 2.3, we briefly discuss the problem of choosing the best regression prediction model. In Section 2.4, we establish the relationship between the problem of pooling data and the problem of choosing the best regression prediction model in the sense that the always-pooled, the never-pooled and the sometimes-pooled estimators correspond respectively to what we call the deleted-model, the full-model and the conditional predictors.

2.2. A general survey of literature concerning preliminary tests and pooling of data.

In many practical problems the statistician is often uncertain of some assumptions required to validate a desired inference procedure. The inference problem may be either testing hypotheses, estimation, or prediction. A lot of statistical literature so far has used the method of preliminary tests to ascertain the assumptions which are most suitable for the inference of major interest. Consequently, the inference of major interest is conditional on the outcome of the preliminary tests of significance.

The testing hypotheses problems using preliminary tests of significance are mainly concerned with analysis of variance. Bozivich, Bancroft and Hartley (1956) and Bancroft (1964) have detailed discussions on the method of using preliminary tests in analysis of variance.

Suppose we are given three mean squares as follows: (i) treatment mean square V_3 with n_3 degrees of freedom, (ii) error mean square V_2 with n_2 degrees of freedom, (iii) doubtful error mean square V_1 with n_1 degrees of freedom. We want to test H_0: no treatment effects. We have two testing procedures as follows: (i) Never-pool procedure: Compare V_3 with V_2 by the F-test as we do in a conventional way. (ii) Sometimes-pool procedure: Perform a preliminary test by comparing V_2 against V_1 by the F-test. If this turns out to be non-significant, then use $V = (n_1V_1 + n_2V_2)/(n_1 + n_2)$ as error for comparison with V_3 in the final F-test. If V_2 is significantly different from V_1, use V_2 in the final F-test.

As an example, suppose we have a nested model

$$(2.1) \qquad Y_{ijk} = \mu + a_i + b_{i(j)} + z_{ij(k)} ,$$

where $i = 1,2,\ldots,I$; $j = 1,2,\ldots,J$; $k = 1,2,\ldots,K$; $a_i \sim N(0, \sigma_a^2)$, $b_{i(j)} \sim (0, \sigma_b^2)$ and $z_{ij(k)} \sim N(0, \sigma_z^2)$. Then the ANOVA table is as follows:

Source	d.f.	M.S.	Exp. M.S.
A	$n_3 = I-1$	V_3	$\sigma_3^2 = \sigma_z^2 + K\sigma_b^2 + JK\sigma_a^2$
B:A	$n_2 = (J-1)I$	V_2	$\sigma_2^2 = \sigma_z^2 + K\sigma_b^2$
Z:AB	$n_1 = (K-1)IJ$	V_1	$\sigma_1^2 = \sigma_z^2$

Our major interest is to test $H_0:\sigma_3^2 = \sigma_2^2$ $(\sigma_a^2 = 0)$. Bust we suspect that $\sigma_b^2 = 0$. So we test $H_0:\sigma_b^2 = 0$ first. If we accept $H_0:\sigma_b^2 = 0$, we pool V_2 with V_1 and use it as new error mean square. If we reject $H_0:\sigma_b^2 = 0$, we don't pool and use V_2 as error mean square in testing $H_0:\sigma_3^2 = \sigma_2^2$ $(\sigma_a^2 = 0)$.

When we use the method of preliminary tests, the testing procedure is as follows: Reject H_0: no treatment effects if

$$(2.2) \qquad \begin{cases} \text{either } V_2/V_1 \geq F_{n_2,n_1}(\alpha_1) & \text{and } V_3/V_2 \geq F_{n_3,n_2}(\alpha_2) \\ \text{or } \quad V_2/V_1 \leq F_{n_2,n_1}(\alpha_1) & \text{and } V_3/V \geq F_{n_3,n_1+n_2}(\alpha_3) , \end{cases}$$

where $V = (n_1V_1 + n_2V_2)/(n_1 + n_2)$. Let P = power of test. Then $P = P_1 + P_2$, where

$$(2.3) \qquad \begin{aligned} P_1 &= P_r\{V_2/V_1 \geq F_{n_2,n_1}(\alpha_1) \text{ and } V_3/V_2 \geq F_{n_3,n_2}(\alpha_2)\} \\ P_2 &= P_r\{V_2/V_1 \leq F_{n_2,n_1}(\alpha_1) \text{ and } V_3/V \geq F_{n_3,n_1+n_2}(\alpha_3)\} . \end{aligned}$$

P turns out to be a function of the degrees of freedom n_1, n_2 and n_3, $\theta_{32} = \sigma_3^2/\sigma_2^2$, $\theta_{21} = \sigma_2^2/\sigma_1^2$, and the levels of significance α_1, α_2 and α_3. Of these 8 parameters, the degrees of freedom n_1, n_2 and n_3 are known; θ_{32} and θ_{21} are generally unknown. And generally we choose $\alpha_2 = \alpha_3 = 0.05$. Only α_1, the level of significance of the preliminary test, is entirely at our disposal. The behavior of P is studied by Bozivich et al (1956). They indicate that for small θ_{21} the sometimes-pool procedure is more powerful than the never-pool procedure, while for large θ_{21} the situation is reversed.

The other inference problem involving preliminary tests of signficance is estimation. Bancroft (1944) has considered the problem of pooling two sample variances based on the preliminary tests. Suppose S_1^2 and S_2^2 are two independent estimators of variances σ_1^2 and σ_2^2 respectively. We assume that $n_1 S_1^2/\sigma_1^2 \sim X^2_{(n_1)}$ and $n_2 S_2^2/\sigma_2^2 \sim X^2_{(n_2)}$. Our main interest is to estimate σ_1^2. We have three procedures to estimate σ_1^2. (i) Never-pool: Use S_1^2 always. (ii) Always-pool: Use $(n_1 S_1^2 + n_2 S_2^2)/(n_1 + n_2)$. (iii) Sometimes-pool: Use the test of significance of S_1^2/S_2^2 as a criterion in making the decision as to whether to pool the two mean squares or not. That is, test S_1^2/S_2^2 by the F-test. If F is non-significant, use $(n_1 S_1^2 + n_2 S_2^2)/(n_1 + n_2)$ as the estimator of σ_1^2. If F is significant, use S_1^2 as the estimator of σ_1^2. The comparison of mean square errors of three procedures is studied numerically by different sets of parameters σ_1^2 and σ_2^2.

The other aspect of estimation problem involving preliminary tests of significance is concerned with the problem of pooling two

sample means from two independent populations (especially normal) with suspected identical means. Mosteller (1948) considers the case of two independent normal populations with the equal known variance σ^2 but unknown means μ_1 and μ_2. We have past data at our disposal, say two random samples of equal size n from two populations respectively. Let $\bar{Y}_1$ and $\bar{Y}_2$ be the two sample means. The main interest is to estimate μ_1. We can use either $\bar{Y}_1$ or the pooled estimator $(\bar{Y}_1 + \bar{Y}_2)/2$ to estimate μ_1. Or we can test $H_0:\mu_1 = \mu_2$ first. If we accept H_0, we use $(\bar{Y}_1 + \bar{Y}_2)/2$ to estimate μ_1; and if we reject H_0, we use $\bar{Y}_1$. The mean square errors of three estimators as functions of $(\mu_1-\mu_2)/(\sigma\sqrt{\pi/2})$ are compared with each other. Mosteller also considers the case that $\mu_1-\mu_2$ can be treated as normally distributed with mean zero and variance $a^2\sigma^2$. In that case he recommends using the maximum likelihood estimator $\hat{\mu}_1 = [\bar{Y}_1(1 + na^2) + \bar{Y}_2]/(2 + na^2)$.

Kale and Bancroft (1967) discuss the problem of pooling two sample means in the discrete data case. We are given two random samples $(Y_{1j}, j = 1,2,\ldots,n_1)$ and $(Y_{2j}, j = 1,2,\ldots,n_2)$ from two discrete distributions having the same mathematical form. If the observed Y are transformed to X, where $X = f(Y)$, the transofmed observations X_{ij} sometimes may be assumed to be independently $N(\mu_i, \sigma^2)$, $i = 1,2$, where σ^2 is known. For example, if $Y \sim$ Poisson (λ), for large λ, $\sqrt{Y}$ is approximately distributed as $N(\sqrt{\lambda}, 1/4)$. If $Y \sim$ Binomial (n,p), for large n,

$\sin^{-1}\sqrt{Y/n}$ is approximately distributed as $N(\sin^{-1}\sqrt{p}, 1/(4n))$. Thus we can reduce the problem of pooling two sample means in the discrete data case to the problem of pooling two sample means in the case of two independent normal distributions with known variances. Given two independent random samples of size n_1 and n_2 from $N_1(\mu_1, \sigma^2)$ and $N_2(\mu_2, \sigma^2)$ respectively, σ^2 being known, the sometimes-pooled estimator of μ_1 is defined as follows:

$$(2.4) \qquad \bar{X}^* = \begin{cases} \bar{X}_1 & , \text{ If } |\bar{X}_1-\bar{X}_2| \geq c_\alpha \sigma_z \\ (n_1\bar{X}_1+n_2\bar{X}_2)/(n_1+n_2), & \text{if } |\bar{X}_1-\bar{X}_2| < c_\alpha \sigma_z, \end{cases}$$

where $\bar{X}_i$ is the sample mean from $N_i(\mu_i, \sigma^2)$, $\sigma_z^2 = \sigma^2(1/n_1 + 2/n_2)$, and c_α is the solution of $1-\Phi(c_\alpha) = \alpha/2$, where Φ denotes the standard normal distribution function. Let $e = \text{M.S.E.}(\bar{X}_1)/\text{M.S.E.}(\bar{X}^*)$. The behavior of e is graphed and studied by Kale and Bancroft (1967).

Han and Bancroft (1968) consider the case of two independent normal populations with same but unknown variances. Let $\bar{Y}_i$ and S_i^2 be the mean and variance of a random sample of size n_i drawn from $N(\mu_i, \sigma^2)$, $i = 1,2$. It is suspected that $\mu_1 = \mu_2$. In this case, the sometimes-pooled estimator for μ_1 is defined as follows:

$$(2.5) \qquad \bar{Y}^* = \begin{cases} \bar{Y}_1 & , \text{ if } |t| \geq t_\alpha \\ (n_1\bar{Y}_1+n_2\bar{Y}_2)/(n_1+n_2), & \text{if } |t| < t_\alpha, \end{cases}$$

where $t = (\bar{Y}_1 - \bar{Y}_2)/(S_p\sqrt{1/n_1+1/n_2})$, $S_p^2 = [(n_1-1)S_1^2 + (n_2-1)S_2^2]/(n_1+n_2-2)$ and t_α is the $(1-\alpha/2)$th quantile of t distribution with $n_1 + n_2 - 2$ degrees of freedom. Let $e = \text{M.S.E.}(\bar{Y}_1)/\text{M.S.E.}(\bar{Y}^*)$.

The behavior of e is graphed and studied. If $|(\mu_2-\mu_1)/\sigma| \leq \sqrt{1/n_1 + 1/n_2}$, the mean square error of the always-pooled estimator is smaller than the mean square error of either the sometimes-pooled or the never-pooled estimator. If $|(\mu_2-\mu_1)/\sigma| > \sqrt{1/n_1 + 1/n_2}$, there is no uniformly best estimator of those studied by Han and Bancroft (1968) in the mean square error sense.

The problem of estimation involving the preliminary tests of significance can be stated briefly as follows: for a given significance level we test the hypothesis that the parameters of two populations are equal; we pool the data if we accept the hypothesis; we don't pool if we reject the hypothesis. For different significance levels we have different pooling rules. Huntsberger (1955) deals with this problem in more generalized terms. A weighted estimator for the parameter is obtained by using weights which are determined by the observed values of the preliminary test statistic. Suppose we have two independent populations with parameters θ_1 and θ_2 respectively. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be the best estimators of θ_1 and θ_2 provided by statistical theory. If $\theta_1 = \theta_2$, a pooled estimator $g(\hat{\theta}_1, \hat{\theta}_2)$ will, in general, provide an estimator for θ_1 which is better in some sense than $\hat{\theta}_1$. Our main interest is to estimate θ_1. Let T be the statistic which the statistical theory indicates will provide the best test of the hypothesis H_0: $\theta_1 = \theta_2$. We can estimate θ_1 by

$$W(T) = f(T)\,\hat{\theta}_1 + [1 - f(T)]g(\hat{\theta}_1, \hat{\theta}_2)\,, \tag{2.6}$$

where $f(T)$ is a function of T only and can be called "weighting function". The choice of $f(T)$ is restricted to the class of single-valued functions of T which are continuous except on a set of measure 0 and which satisfy the following conditions: (i) $0 \leq f(T) \leq 1$, for all T, (ii) $f(-T) = f(T)$. If $f(T)$ is defined as

$$f(T) = \begin{cases} 0, & T \in A_\alpha \\ 1, & T \in A_\alpha^c \end{cases},$$

where A_α and A_α^c are the acceptance and rejection regions for the test of $H_0: \theta_1 = \theta_2$ with significance level = α, then $W(T)$ reduces to the "sometimes-pooled" estimator which we have already mentioned.

Huntsberger (1955) points out two important facts which are worth our attention. They are

(2.7)

(i) among the class of weighting functions the only unbiased weighting function is $f(T) = 1$,

(ii) there does not exist an estimator which has the uniform minimum mean square error.

Therefore, "sometimes-pooled" estimator using the preliminary test of significance is always a biased estimator. And we can not find a "sometimes-pooled" estimator with a given significance level to be uniformly better (in the sense of mean square error) than other "sometimes-pooled" estimators with different significance levels.

2.3. The problem of choosing a regression prediction model.

In standard linear regression models where the distribution of a dependent variable Y depends on several independent variables X_i it is well known that the mean square error of a predicted future observation may be smaller when it is based on a "deleted model" (where some of the X_i deleted) than when the full-model predictor is used.

Suppose we have the following standard linear regression model:

$$(3.1) \qquad \underset{\sim}{Y} = X_1 \underset{\sim}{B}_1 + X_2 \underset{\sim}{B}_2 + \underset{\sim}{e},$$

where X_1 and X_2 are known, $E(\underset{\sim}{e}) = \underset{\sim}{0}$ and $Var(\underset{\sim}{e}) = \sigma^2 I$. The least square estimator for $\underset{\sim}{B}_1$ and $\underset{\sim}{B}_2$ in (3.1) are as follows:

$$(3.2) \qquad \begin{bmatrix} \hat{\underset{\sim}{B}}_1 \\ \hat{\underset{\sim}{B}}_2 \end{bmatrix} = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} \underset{\sim}{Y} .$$

Suppose we delete the X_2 independent variable. Then

$$(3.3) \qquad \underset{\sim}{Y} = X_1 \underset{\sim}{B}_1 + \underset{\sim}{e} .$$

The least square estimator for $\underset{\sim}{B}_1$ in (3.3) is

$$(3.4) \qquad \hat{\hat{\underset{\sim}{B}}}_1 = (X_1'X_1)^{-1} X_1' \underset{\sim}{Y} .$$

Suppose we have a future observation Y_0 such that

$$(3.5) \qquad Y_0 = \underset{\sim}{X}_{10}' \underset{\sim}{B}_1 + \underset{\sim}{X}_{20}' \underset{\sim}{B}_2 + e_0 ,$$

where $\underset{\sim}{X}_{i0}'$, $i = 1,2$, are known, $E(e_0) = 0$, $Var(e_0) = \sigma^2$ and e_0 is independent of $\underset{\sim}{e}$. We want to predict Y_0. We can use

(3.6) $\hat{Y}_0 = \underset{\sim}{X}_{10}' \hat{\underset{\sim}{B}}_1 + \underset{\sim}{X}_{20}' \hat{\underset{\sim}{B}}_2$

to predict Y_0. We will call $\hat{Y}_0$ the full-model predictor. Or we can use

(3.7) $\hat{\hat{Y}}_0 = \underset{\sim}{X}_{10}' \hat{\hat{\underset{\sim}{B}}}_1$

to predict Y_0. We will call $\hat{\hat{Y}}_0$ the deleted-model predictor.

It is a well known fact that sometimes $E(\hat{\hat{Y}}_0 - Y_0)^2$ is less than $E(\hat{Y}_0 - Y_0)^2$. A necessary and sufficient condition for which $E(\hat{\hat{Y}}_0 - Y_0)^2 < E(\hat{Y}_0 - Y_0)^2$ is given by Anderson, Allen and Cady (1972). A special case was given earlier by Schneider (1970) for the model $Y_i = \alpha + \beta X_i + e_i$.

The equations (12) up to (15) of Toro-Vizcarrondo and Wallace (1968) give a necessary and sufficient condition for which the mean square error of any linear combination of restricted estimators is less than the mean square error of that of unrestricted estimators. The deleted-model predictor $\hat{\hat{Y}}_0$ is a linear combination of "restricted" estimators; while the full-model predictor $\hat{Y}_0$ is a linear combination of unrestricted estimators. It is easy to see that $E(\hat{\hat{Y}}_0 - Y_0)^2$ is equal to σ^2 plus the mean square error of its corresponding linear combination of "restricted" estimators. Similarly, $E(\hat{Y}_0 - Y_0)^2$ is equal to σ^2 plus the mean square error of its corresponding linear combination of unrestricted estimators. Thus the condition of Toro-Vizcarrondo et al implies the condition of Anderson et al, but is not equivalent to it.

In order to have the present thesis self contained and to have some formulas to refer to later we rederive necessary and sufficient conditions essentially equivalent to those of Anderson et al (1972).

Lemma 3.1.

$$(3.8)\qquad [\underset{\sim}{X}_{10}' : \underset{\sim}{X}_{20}'] \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} \underset{\sim}{X}_{10} \\ \underset{\sim}{X}_{20} \end{bmatrix} = \underset{\sim}{X}_{10}'(X_1'X_1)^{-1}\,\underset{\sim}{X}_{10} + \underset{\sim}{Z}_0'\,W^{-1}\,\underset{\sim}{Z}_0,$$

where

$$(3.9)\qquad \underset{\sim}{Z}_0 = \underset{\sim}{X}_{20} - X_2'X_1(X_1'X_1)^{-1}\,\underset{\sim}{X}_{10}\,,$$

$$(3.10)\qquad W = (X_2'X_2) - X_2'X_1(X_1'X_1)^{-1}\,X_1'X_2\,.$$

Proof: It is well known that

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} =$$

$$(3.11)\qquad \begin{bmatrix} (X_1'X_1)^{-1}+(X_1'X_1)^{-1}X_1'X_2W^{-1}X_2'X_1(X_1'X_1)^{-1} & -(X_1'X_1)^{-1}X_1'X_2W^{-1} \\ -W^{-1}X_2'X_1(X_1'X_1)^{-1} & W^{-1} \end{bmatrix}.$$

Use (3.11) and do some straightforward algebra, we have the result. Q.E.D.

Lemma 3.2.

$$(3.12)\qquad [\underset{\sim}{X}_{10}' : \underset{\sim}{X}_{20}'] \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} \underset{\sim}{X}_{10} \\ \underset{\sim}{X}_{20} \end{bmatrix} \geq \underset{\sim}{X}_{10}'(X_1'X_1)^{-1}\,\underset{\sim}{X}_{10}.$$

Proof: It follows directly from (3.8).

Lemma 3.3.

(3.13) $\quad \mathrm{Var}(\hat{Y}_0) = \sigma^2 \underset{\sim}{X}_{10}'(X_1'X_1)^{-1} \underset{\sim}{X}_{10} + \sigma^2 \underset{\sim}{Z}_0' W^{-1} \underset{\sim}{Z}_0 .$

(3.14) $\quad \mathrm{Var}(\hat{\hat{Y}}_0) = \sigma^2 \underset{\sim}{X}_{10}'(X_1'X_1)^{-1} \underset{\sim}{X}_{10} .$

Proof: From (3.2), (3.6) and (3.8), we have (3.13). From (3.4) and (3.7), we have (3.14). Q.E.D.

Lemma 3.4.

(3.15) $\quad \mathrm{Var}(\hat{\hat{Y}}_0) \leq \mathrm{Var}(\hat{Y}_0) .$

Proof: It follows directly from (3.13) and (3.14).

Lemma 3.5.

(3.16) $\quad E(\hat{\hat{Y}}_0 - Y_0)^2 = \sigma^2 + \sigma^2 \underset{\sim}{X}_{10}'(X_1'X_1)^{-1} \underset{\sim}{X}_{10} + \sigma^2 \underset{\sim}{Z}_0' W^{-1} \underset{\sim}{Z}_0 .$

(3.17) $\quad E(\hat{\hat{Y}}_0 - Y_0)^2 = \sigma^2 + \sigma^2 \underset{\sim}{X}_{10}'(X_1'X_1)^{-1} \underset{\sim}{X}_{10} + (\underset{\sim}{Z}_0' \underset{\sim}{B}_2)^2 .$

Proof: (3.16) follows from (3.2), (3.5), (3.6) and (3.8). (3.17) follows from (3.4), (3.5), (3.7) and (3.9). Q.E.D.

Theorem 3.1. (This result follows immediately from Equations (3.8) and (3.10) of Anderson et al (1972).)

(3.18) $\quad E(\hat{\hat{Y}}_0 - Y_0)^2 \leq E(\hat{Y}_0 - Y_0)^2 \quad \text{iff} \quad (\underset{\sim}{Z}_0' \underset{\sim}{B}_2)^2 \leq \sigma^2 \underset{\sim}{Z}_0' W^{-1} \underset{\sim}{Z}_0 .$

Proof: (3.18) follows directly from (3.16) and (3.17). Q.E.D.

Corollary 3.1. (Schneider (1970)). Suppose we have the special model: $Y_i = \alpha + \beta X_i + e_i$, $i = 1,2,\ldots,n$. Let the deleted model

be $Y_i = \alpha + e_i$. Then

$$(3.19) \quad E(\hat{Y}_0 - Y_0)^2 \leq E(\hat{Y}_0 - Y_0)^2 \text{ iff } \beta^2 \leq \sigma^2 / \Sigma(X_i - \bar{X})^2 .$$

Proof: It follows from (3.18).

The main problem is to choose the best linear regression prediction model. Anderson et al (1972) indicate the following decision rule to choose between the full-model predictor (3.6) and the deleted-model predictor (3.7). Let

$$(3.20) \quad W = \frac{(\underset{\sim}{Z}_0' \hat{\underset{\sim}{B}})^2 (n-k)}{(\underset{\sim}{Z}_0' W^{-1} \underset{\sim}{Z}_0)(\text{S.S. E.})} ,$$

where S.S.E. is sum of squares for error in the usual sense. Under normality assumption, $W \sim F(1, n-k; \lambda)$, where

$$(3.21) \quad \lambda = \frac{(\underset{\sim}{Z}_0' \underset{\sim}{B})^2}{2(\underset{\sim}{Z}_0' W^{-1} \underset{\sim}{Z}_0)\sigma^2} .$$

It follows that (3.18) is equivalent to

$$(3.22) \quad E(\hat{Y}_0 - Y_0)^2 \leq E(\hat{Y}_0 - Y_0)^2 \text{ iff } \lambda \leq \tfrac{1}{2} .$$

We test $H_0: \lambda \leq \frac{1}{2}$ against $H_1: \lambda > \frac{1}{2}$. The test is given by

Accept H_0 if $W \leq W_\alpha$

Reject H_0 if $W > W_\alpha$,

where W_α is the $(1-\alpha)$th quantile of $F(1, n-k; \frac{1}{2})$. Then we have the following decision rule:

(3.23) $\begin{cases} \text{Use the deleted-model predictor if } H_0 \text{ is accepted.} \\ \text{Use the full-model predictor if } H_0 \text{ is rejected.} \end{cases}$

The decision rule here is conditional upon our preliminary test of $H_0: \lambda \leq \frac{1}{2}$. However, the performance of the conditional predictor based on this decision rule apparently has not been studied yet. The comparison of two or more deleted models directly is more difficult than comparing a deleted model to a full model. If we have k independent variables, then we have 2^k-1 deleted models to consider. Anderson et al (1972) suggest that we choose the deleted model having the smallest value of $(\underset{\sim}{Z}_0' \hat{\underset{\sim}{B}}_2)^2-[(\underset{\sim}{Z}_0'W^{-1}\underset{\sim}{Z}_0)(\text{S.S.E.})]/(n-k)$. But this rule is based on intuition. The mathematical properties of this rule apparently have not been studied.

2.4. The relationship between the problem of pooling data and the problem of choosing a regression prediction model.

Suppose we have n_1 observations $(Y_{1j}, j = 1,2,\ldots,n_1)$ from $N(\mu_1, \sigma^2)$ and n_2 observations $(Y_{2j}, j = 1,2,\ldots,n_2)$ from $N(\mu_2, \sigma^2)$, σ^2 being unknown and two populations being independent. Our main interest is to estimate μ_1. We have the following two estimators for μ_1:

(4.1) (i) Never-pooled estimator $\hat{\mu}_{1N} = \sum_j Y_{1j}/n_1$

(ii) Always-pooled estimator $\hat{\mu}_{1A} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}/(n_1+n_2)$.

We want to know which estimator is best. We attempt to discuss this problem in the framework of linear regression models and to find the

relationship between the problem of pooling data and the problem of choosing the best linear regression prediction model. Let

$$(4.2)\qquad \begin{aligned} Y_{1j} &= Y_j, \quad j = 1,2,\ldots,n_1 \\ Y_{2j} &= Y_{n_1+j}, \quad j = 1,2,\ldots,n_2 . \end{aligned}$$

Let

$$(4.3)\qquad \alpha = \mu_2, \quad \beta = \mu_1 - \mu_2 .$$

Then

$$(4.4)\qquad \begin{aligned} Y_j &\sim N(\alpha + \beta, \sigma^2), \quad j = 1,2,\ldots,n_1 \\ Y_j &\sim N(\alpha, \sigma^2), \quad j = n_1+1,\ldots,n_1+n_2 . \end{aligned}$$

Or equivalently,

$$(4.5)\qquad Y_j = \alpha + \beta X_j + e_j, \quad j = 1,2,\ldots,n_1,n_1+1,\ldots,n_1+n_2 ,$$

where

$$(4.6)\qquad X_j = \begin{cases} 1, & j = 1,2,\ldots,n_1 \\ 0, & j = n_1+1,\ldots,n_1+n_2 , \end{cases}$$

e_j's are independently distributed as $N(0, \sigma^2)$.

Suppose we delete the X_j variable. Then

$$(4.7)\qquad Y_j = \alpha + e_j, \quad j = 1,2,\ldots,n_1+n_2 .$$

Let

(4.8) $\hat{\alpha}$ and $\hat{\beta}$ be the least square estimators for α and β in (4.5)

$\hat{\hat{\alpha}}$ be the least square estimator for α in (4.7) .

Then we have the following result.

Lemma 4.1.

$$(4.9)\qquad \begin{aligned} \hat{\mu}_{1N} &= \hat{\alpha} + \hat{\beta} = \sum_{j=1}^{n_1} Y_j/n_1 \\ \hat{\mu}_{1A} &= \hat{\hat{\alpha}} = \sum_{j=1}^{n_1+n_2} Y_j/(n_1+n_2) . \end{aligned}$$

Proof: (4.9) follows from (4.2), (4.5), (4.6), (4.7) and (4.8). Q.E.D.

Suppose Y_0 is a future observation and $X_0 = 1$ such that

(4.10) $\quad Y_0 = \alpha + \beta + e_0 ,$

where e_0 is independent of e_j and $e_0 \sim N(0, \sigma^2)$. Then the full-model and the deleted-model predictors are respectively as follows:

(4.11) $\quad \hat{Y}_0 = \hat{\alpha} + \hat{\beta}, \quad \hat{\hat{Y}}_0 = \hat{\hat{\alpha}} .$

We have the following result.

Theorem 4.1.

(4.12) $\quad E(\hat{\hat{Y}}_0 - Y_0)^2 < E(\hat{Y}_0 - Y_0)^2 \Leftrightarrow E(\hat{\mu}_{1A} - \mu_1)^2 < E(\hat{\mu}_{1N} - \mu_1)^2$

$$\Leftrightarrow \beta^2 < \sigma^2 / \textstyle\sum_j (X_j - \bar{X})^2 .$$

Proof: It is easy to see that

(4.13) $\quad E(\hat{\hat{Y}}_0 - Y_0)^2 = \sigma^2 + E(\hat{\hat{\alpha}} - \alpha - \beta)^2 = \sigma^2 + E(\hat{\mu}_{1A} - \mu_1)^2$

$$E(\hat{Y}_0 - Y_0)^2 = \sigma^2 + E(\hat{\alpha} + \hat{\beta} - \alpha - \beta)^2 = \sigma^2 + E(\hat{\mu}_{1N} - \mu_1)^2 .$$

(4.12) follows directly from (3.19) and (4.13). Q.E.D.

Remark: (4.12) tells us that the set of parameters for which the mean square error of the deleted-model predictor is less than that of the full-model predictor coincides with those parameters for which the mean square error of the always-pooled estimator is less than that of the never-pooled estimator. In this sense the always-pooled and the never-pooled estimators coorespond respectively to the deleted-model and the full-model predictors.

Next, we would like to explore the relationship between the sometimes-pooled estimator $\hat{\mu}_{1s}$ and the conditional predictor $\hat{Y}_0^{(s)}$. For any test of hypothesis H_0 we define $\hat{\mu}_{1s}$ as follows:

$$\hat{\mu}_{1s} = \begin{cases} \hat{\mu}_{1N}, & \text{if reject } H_0: \mu_1 = \mu_2 \\ \hat{\mu}_{1A}, & \text{if accept } H_0: \mu_1 = \mu_2 . \end{cases}$$

By (4.3), $H_0: \mu_1 = \mu_2$ is equivalent to $H_0: \beta = 0$. Under $H_0: \beta = 0$,

$$(4.14) \qquad \hat{\beta} = \bar{Y}_1 - \bar{Y}_2 \sim N(0, \sigma^2(1/n_1 + 1/n_2)) ,$$

where

$$(4.15) \qquad \bar{Y}_1 = \sum_{j=1}^{n_1} Y_j/n_1, \quad \bar{Y}_2 = \sum_{j=n_1+1}^{n_1+n_2} Y_j/n_2 .$$

Let

$$(4.16) \qquad \hat{\sigma}^2 = \sum_{j=1}^{n_1+n_2} (Y_j - \hat{\alpha} - \hat{\beta}X_j)^2/(n_1+n_2-2)$$

$$= [\sum_{j=1}^{n_1}(Y_j - \bar{Y}_1)^2 + \sum_{j=n_1+1}^{n_1+n_2}(Y_j - \bar{Y}_2)^2]/(n_1 + n_2 - 2).$$

Then under $H_0: \beta = 0$,

$$(4.17) \qquad t = \frac{\bar{Y}_1 - \bar{Y}_2}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} = \frac{\hat{\beta}}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} \sim t(n_1+n_2-2) .$$

We define $\hat{\mu}_{1s}$ formally as follows:

$$(4.18) \qquad \hat{\mu}_{1s} = \begin{cases} \hat{\mu}_{1N}, & \text{if } |t| > t_\alpha \\ \hat{\mu}_{1A}, & \text{if } |t| \le t_\alpha , \end{cases}$$

where t_α is the $(1-\alpha/2)$th quantile of $t(n_1 + n_2 - 2)$. Then $\hat{\mu}_{1s}$ is exactly the one being studied by Han and Bancroft (1968).

For this particular linear regression model (as indicated in (4.5)), (3.20) is equivalent to

$$(4.19) \qquad W = \frac{\hat{\beta}^2}{\hat{\sigma}^2 / \sum_{j=1}^{n_1+n_2} (X_j - \bar{X})^2} \sim F(1, n_1+n_2-2; \lambda)$$

where

$$(4.20) \qquad \lambda = \frac{\beta^2}{2\sigma^2 / \sum_{j=1}^{n_1+n_2} (X_j - \bar{X})^2} .$$

Following what Anderson et al (1972) suggest (as indicated in (3.23)), we define the conditional predictor $\hat{Y}_0^{(s)}$ as follows:

$$(4.21) \qquad \hat{Y}_0^{(s)} = \begin{cases} \hat{Y}_0, & \text{if } W > W_{\alpha'} \\ \overset{*}{Y}_0, & \text{if } W \le W_{\alpha'} , \end{cases}$$

where $W_{\alpha'}$ is the $(1-\alpha')$th quantile of $F(1, n_1+n_2-2; 1/2)$. Then we have the following result.

Theorem 4.2. If we choose α and α' properly such that $t_\alpha^2 = W_{\alpha'}$, then

$$(4.22) \qquad E(\hat{Y}_0^{(s)} - Y_0)^2 = \sigma^2 + E(\hat{\mu}_{1s} - \mu_1)^2 .$$

Proof: (4.18) is equivalent to

$$(4.23) \qquad \hat{\mu}_{1s} = \begin{cases} \hat{\mu}_{1N}, & \text{if } t^2 > t_\alpha^2 \\ \hat{\mu}_{1A}, & \text{If } t^2 \le t_\alpha^2 . \end{cases}$$

We note that in this particular model (4.5),

$$(4.24) \qquad 1 / \sum_{j=1}^{n_1+n_2} (X_j - \bar{X})^2 = 1/n_1 + 1/n_2 .$$

It follows that $W = t^2$. Suppose we choose α and α' properly such that $t_\alpha^2 = W_{\alpha'}$. Let us consider

$$(4.25)\quad E(\hat{Y}_0^{(s)} - Y_0)^2 = E[(\hat{Y}_0 - Y_0)^2 | W > W_{\alpha'}]P(W > W_{\alpha'}) + E[(\hat{\hat{Y}}_0 - Y_0)^2 | W \le W_{\alpha'}]P(W \le W_{\alpha'}) .$$

But

$$(4.26)\quad E[(\hat{Y}_0 - Y_0)^2 | W > W_{\alpha'}] = E[(\hat{Y}_0 - Y_0)^2 | t^2 > t_\alpha^2]$$

$$= E[(\hat{\alpha} + \hat{\beta} - \alpha - \beta - e_0)^2 | t^2 > t_\alpha^2]$$

$$= E[(\hat{\alpha} + \hat{\beta} - \alpha - \beta)^2 + e_0^2 - 2e_0(\hat{\alpha} + \hat{\beta} - \alpha - \beta) | t^2 > t_\alpha^2] .$$

Since e_0 is independent of $\hat{\alpha}$, $\hat{\beta}$ and t, (4.26) is equivalent to

$$(4.27)\quad E[(\hat{Y}_0 - Y_0)^2 | W > W_{\alpha'}] = \sigma^2 + E[(\hat{\mu}_{1N} - \mu_1)^2 | t^2 > t_\alpha^2] .$$

Similarly,

$$(4.28)\quad E[(\hat{\hat{Y}}_0 - Y_0)^2 | W \le W_{\alpha'}] = \sigma^2 + E[(\hat{\mu}_{1A} - \mu_1)^2 | t^2 \le t_\alpha^2] .$$

From (4.25), (4.27) and (4.28), we have

$$(4.29)\quad E(\hat{Y}_0^{(s)} - Y_0)^2 = \{\sigma^2 + E[(\hat{\mu}_{1N} - \mu_1)^2 | t^2 > t_\alpha^2]\}P(t^2 > t_\alpha^2) + \{\sigma^2 + E[(\hat{\mu}_{1A} - \mu_1)^2 | t^2 \le t_2^2]\}P(t^2 \le t_\alpha^2)$$

$$= \sigma^2 + E(\hat{\mu}_{1s} - \mu_1)^2 . \quad \text{Q.E.D.}$$

Remarks: (i) Toro-Vizcarrondo et al (1968) point out that $\alpha = 0.01$ is equivalent to $\alpha' = 0.05$; $\alpha = 0.03$ is equivalent to $\alpha' = 0.10$. (ii) Once we know the behavior of $E(\hat{\mu}_{1s} - \mu_1)^2$, we know the behavior of $E(\hat{Y}_0^{(s)} - Y_0)^2$. The behavior of $E(\hat{\mu}_{1s} - \mu_1)^2$ has been studied in detail by Han and Bancroft (1968).

CHAPTER 3

THE PROBLEM OF POOLING DATA IN THE TWO POPULATIONS CASE (USING A MEAN SQUARE ERROR CRITERION)

3.1. Introduction.

In Chapter 2, we have mentioned that under the situation of availability of past data, "ignoring stratifications" is equivalent to "pooling of data". We have indicated that in the literature the method of preliminary tests of significance is applied to the problem of pooling data. The idea behind this is that we should pool the data if the parameters of two populations are equal, since pooling of data will provide additional information. Since we don't know if the parameters of two populations are equal, we make a preliminary test. If we accept the hypothesis, then we pool data; otherwise, we don't pool.

The next question is whether the pooling rule based on the preliminary tests of significance is the best pooling rule. In this chapter, we introduce another pooling rule. Let us call the set of parameters for which the mean square error of the always-pooled estimator is less than that of the never-pooled the "pooling region". If parameters fall in the "pooling region" then it is best to pool the data. Since parameters are unknown, we replace them by estimators. We call the sometimes-pooled estimator based on this pooling rule "the sometimes-pooled estimator based on the estimated mean square error".

The intuitive idea behind the pooling rule based on the preliminary tests of significance is that we pool the data only if

parameters of two populations are equal. The intuitive idea behind the pooling rule based on estimated mean square error is that we pool the data when parameters of two populations fall in the "pooling region". Whenever parameters of two populations are equal, they will automatically fall in the "pooling region".

In Section 3.2, we study in detail the "pooling region" in the case of two binomial populations. In Section 3.3, in the case of two binomial populations, we propose a sometimes-pooled estimator based on a linearly approximated mean square error. We carry out a numerical study to compare the performance of this estimator to that of sometimes-pooled estimator based on preliminary tests in the Kale-Bancroft (1967) sense. The numerical results indicate that when the difference of two parameters is large, the sometimes-pooled estimator based on the linearly approximated mean square error has a better performance. In Section 3.4, we briefly discuss the normal and the Poisson cases. In the normal case the sometimes-pooled estimator based on the estimated mean square error is just a special case of the one based on the preliminary tests of significance when we choose the right significance level.

3.2. Never-pooled and always-pooled estimators for two binomial populations.

Suppose we have two binomial populations, say π_1 (with parameter p) and π_2 (with parameter $p + \tau$). It is clear that

(2.1) $\quad 0 < p < 1 \quad \text{and} \quad -p < \tau < 1 - p$.

We can consider p (or $p + \tau$) as the probability of failure in each

Bernoulli trial. Suppose we have the following past data available:

(2.2) (i) X_1 defectives out of n_1 observations from π_1,
(ii) X_2 defectives out of n_2 observations from π_2,
X_1 and X_2 are independent.

Our problem is to estimate p. We have two estimators for p as follows:

(2.3) (i) Never-pooled estimator $\hat{p}_N = X_1/n_1$,
(ii) Always-pooled estimator $\hat{p}_A = (X_1+ X_2)/(n_1+ n_2)$.

We want to know when $\hat{p}_A$ is better than $\hat{p}_N$ in the mean square error sense, and vice versa.

Let

(2.4) $R = \{(p,\tau)|0 < p < 1 \text{ and } -p < \tau < 1-p\}$

be the allowable region of (p,τ). Then the boundary of R is

(2.5) $\partial R = \{(p,\tau)|p = 0 \text{ and } 0 \le \tau \le 1\}\cup\{(p,\tau)|0 < p < 1 \text{ and } \tau = 1-p\}\cup$
$\{(p,\tau)|0 < p < 1 \text{ and } \tau = -p\}\cup\{(p,\tau)|p = 1 \text{ and } -1 \le \tau \le 0\}$.

Let

(2.6) $D(p,\tau) = E(\hat{p}_A - p)^2 - E(\hat{p}_N - p)^2$
$= ap^2 + bp\tau + c\tau^2 + dp + e\tau$,

where $a = n_2/[n_1(n_1+n_2)]$, $b = -2n_2/(n_1+n_2)^2$, $c = n_2(n_2-1)/(n_1+n_2)^2$, $d = -a$ and $e = -b/2$. Let

(2.7) $D^*(p,\tau) = \{(p,\tau)|D(p,\tau) = 0\}$.

Then we have the following results:

Lemma 2.1.

(2.8) When $n_2 \ge 2$, $D^*(p,\tau)$ is an ellipse.

Proof: (2.8) follows directly from (2.7).

Therefore,

(2.9) We always assume $n_2 \geq 2$ in the following discussions.

Lemma 2.2.

(2.10) (i) $D(p,\tau)$ is minimum at $p = \frac{1}{2}$ and $\tau = 0$.
(ii) $(\frac{1}{2},0) \in R$.

Proof: See Appendix B.

Remark: When $p = \frac{1}{2}$ and $\tau = 0$, we get the maximum gain by using the always-pooled estimator as compared to the never-pooled estimator.

Lemma 2.3. Consider $(p,\tau) \in \{(p,\tau)|p = 0 \text{ and } 0 \leq \tau \leq 1\}$. Then

(2.11) (i) $D(p,\tau) = 0$, if $p = 0$ and $\tau = 0$,
(ii) $D(p,\tau) > 0$, if $(p,\tau) \in \{(p,\tau)|p = 0 \text{ and } 0 < \tau \leq 1\}$.

Proof: See Appendix B.

Lemma 2.4. Consider $(p,\tau) \in \{(p,\tau)|0 < p < 1 \text{ and } \tau = 1-p\}$. Then

(2.12) (i) $D(p,\tau) > 0$, if $(p,\tau) \in \{(p,\tau)|0 < p < p^* \text{ and } \tau = 1-p\}$,
(ii) $D(p,\tau) = 0$, if $p = p^*$ and $\tau = 1 - p^*$,
(iii) $D(p,\tau) < 0$, if $(p,\tau) \in \{(p,\tau)|p^* < p < 1 \text{ and } \tau = 1-p\}$,

where

(2.13) $p^* = n_1 n_2/(n_2+2n_1+n_1 n_2)$.

Proof: See Appendix B.

Lemma 2.5. Consider $(p,\tau) \in \{(p,\tau) | 0 < p < 1 \text{ and } \tau = -p\}$. Then

$$\text{(2.14)} \quad \begin{aligned} &\text{(i)} \quad D(p,\tau) < 0, \text{ if } (p,\tau) \in \{(p,\tau) | 0 < p < p^{**} \text{ and } \tau = -p\}, \\ &\text{(ii)} \quad D(p,\tau) = 0, \text{ if } p = p^{**} \text{ and } \tau = -p^{**}, \\ &\text{(iii)} \quad D(p,\tau) > 0, \text{ if } (p,\tau) \in \{(p,\tau) | p^{**} < p < 1 \text{ and } \tau = -p\}, \end{aligned}$$

where

$$\text{(2.15)} \quad p^{**} = 1 - p^{*} = (n_2 + 2n_1)/(n_2 + 2n_1 + n_1 n_2) .$$

Proof: See Appendix B.

Lemma 2.6. Consider $(p,\tau) \in \{(p,\tau) | p = 1 \text{ and } -1 \le \tau \le 0\}$. Then

$$\text{(2.16)} \quad \begin{aligned} &\text{(i)} \quad D(p,\tau) = 0, \text{ if } p = 1 \text{ and } \tau = 0, \\ &\text{(ii)} \quad D(p,\tau) > 0, \text{ if } (p,\tau) \in \{(p,\tau) | p = 1 \text{ and } -1 \le \tau < 0\}. \end{aligned}$$

Proof: See Appendix B.

Let

$$\text{(2.17)} \quad D^{**}(p,\tau) = \{(p,\tau) | D(p,\tau) < 0\}, \; D^{***}(p,\tau) = \{(p,\tau) | D(p,\tau) > 0\}$$
$$R_1 = R \cap D^{**}(p,\tau) \text{ and } R_2 = R \cap D^{***}(p,\tau) .$$

From (2.4) up to (2.17), we have Figures 2.1 and 2.2. The shaded region is R_2. The unshaded region is R_1.

Lemma 2.7.

$$\text{(2.18)} \quad \max_{(p,\tau) \in R \cup \partial R} D(p,\tau) = D(0,1) = D(1,-1) = \frac{n_2^2}{(n_1 + n_2)^2} .$$

Proof: Since $D^{*}(p,\tau)$ is an ellipse, it is clear that $D(p,\tau) \uparrow \infty$ as $p \uparrow \infty$ and $\tau \downarrow -\infty$ such that $\tau = -p$. Similarly, $D(p,\tau) \uparrow \infty$ as $p \downarrow -\infty$ and $\tau \uparrow \infty$ such that $\tau = 1-p$. If we restrict $D(p,\tau)$ to $(p,\tau) \in R \cup \partial R$, it is clear that $D(p,\tau)$ is maximum at $(0,1)$ or $(1, -1)$. It is easy to see that $D(0,1) = D(1,-1) = n_2^2/(n_1 + n_2)^2$. Q.E.D.

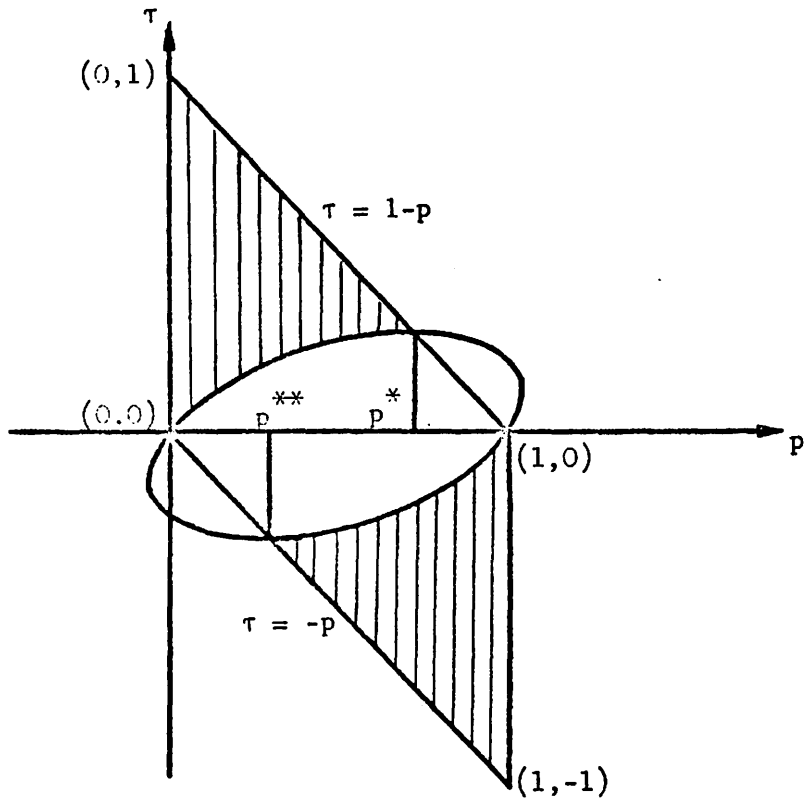


Figure 2.1 $(p^{**} < \frac{1}{2} < p^{*})$

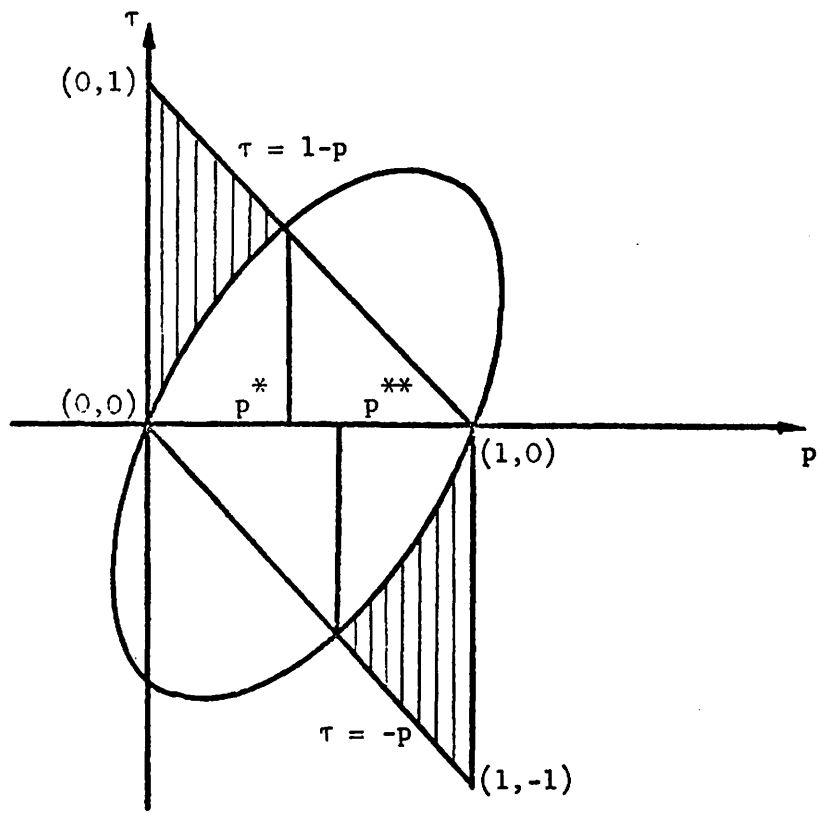


Figure 2.2 $(p^{*} < \frac{1}{2} < p^{**})$

Remark: When either (i) p is close to 0 and τ is close to 1 or (ii) p is close to 1 and τ is close to -1, we will get maximal loss by using the always-pooled estimator.

Lemma 2.8. If $\tau = 0$, then $D(p,\tau) < 0$ for all $0 < p < 1$. (See Figure 2.1 and Figure 2.2.)

Proof: It follows directly from (2.6).

3.3. Sometimes-pooled estimators for two binomial populations.

In Section 3.2, we have shown that sometimes we can gain by using the always-pooled estimator especially when $\tau = 0$. As we have mentioned earlier, the ideas of preliminary tests of significance give us a pooling rule. We carry out a preliminary test of the hypothesis that the parameters of two populations are equal. If we accept the hypothesis, then we use the always-pooled estimator; otherwise, we use the never-pooled estimator. As we have mentioned in Chapter 2, Kale and Bancroft (1967) have given us a pooling rule based on the preliminary test in the case of two binomial populations. They use the arcsine square root transformation and normal approximations, and thus transform the problem to the problem of pooling two sample means in the case of two independent normal distributions with known variances.

In Section 3.2, in the case of two binomial populations we have studied in detail the "pooling region ", that is the region of (p,τ) for which the always-pooled estimator is better than the never-pooled estimator. An intuitively appealing pooling rule will be that we use the always-pooled estimator if $(\hat{p}, \hat{\tau})$ falls in the "pooling region", where $(\hat{p}, \hat{\tau})$ is the standard estimator of (p,τ). However, the "pooling region" is quite complex mathematically. We consider a "linearly approximated pooling region". We consider a pooling rule as follows: use the always-pooled estimator if $(\hat{p}, \hat{\tau})$ falls in the "linearly approximated pooling region". We shall call the estimator based on this pooling rule "the sometimes-pooled estimator based on a linearly approximated mean square error". A numerical study to compare the performance of this estimator to that of the sometimes-pooled estimator in the Kale - Bancroft (1967) sense is presented at the end of this section.

Before we go into details, the following lemmas will be needed for later discussions. Let ϕ and Φ stand for standard normal p.d.f. and c.d.f. respectively.

Lemma 3.1.

$$\text{(i)} \quad \int_a^b x^2 \, \phi(x) \, dx = a\,\phi(a) - b\,\phi(b) + \Phi(b) - \Phi(a)$$

$$\text{(3.1)} \quad \text{(ii)} \quad \int_a^b x \, \phi(x) \, dx = \phi(a) - \phi(b)$$

$$\text{(iii)} \quad \int_a^b \phi(x) \, dx = \Phi(b) - \Phi(a) \, .$$

Proof: By integration by parts and straightforward algebra, we have (3.1). Q.E.D.

Lemma 3.2. Let $X \sim N(\mu, \sigma^2)$ and $c > 0$. Let $n(x)$ be the p.d.f. of X.

$$\text{(i)} \quad \int_0^c x^2\, n(x)dx$$

$$= \sigma^2[c_a\phi(c_a) - c_b\phi(c_b) + \Phi(c_b) - \Phi(c_a)] + \mu^2[\Phi(c_b) - \Phi(c_a)]$$

$$+ 2\mu\sigma[\phi(c_a) - \phi(c_b)] \,, \tag{3.2}$$

$$\text{(ii)} \quad \int_0^c x\, n(x)dx = \sigma[\phi(c_a) - \phi(c_b)] + \mu[\Phi(c_b) - \Phi(c_a)] \,,$$

$$\text{(iii)} \quad \int_0^c n(x)dx = \Phi(c_b) - \Phi(c_a) \,,$$

where $c_a = -\mu/\sigma$ and $c_b = (c-\mu)/\sigma$.

Proof: (3.2) follows from (3.1).

Lemma 3.3. Suppose

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right) .$$

Let

$A = \{(x_1,x_2) \mid 0 \le x_1 \le c, \ -\infty < x_2 < \infty\}$ and $f(x_1,x_2)$ be the p.d.f. of (X_1,X_2). Then

$$\text{(i)} \quad \iint_A x_1^2 f(x_1,x_2)dx_1dx_2$$

$$= \sigma_1^2[c_a\phi(c_a) - c_b\phi(c_b) + \Phi(c_b) - \Phi(c_a)] + \mu_1^2[\Phi(c_b)-\Phi(c_a)]$$

$$+ 2\mu_1\sigma_1[\phi(c_a) - \phi(c_b)] \,,$$

$$\text{(ii)} \quad \iint_A x_2^2 f(x_1,x_2)dx_1dx_2$$

$$= [\sigma_2^2 + \mu_2^2][\Phi(c_b) - \Phi(c_a)] + 2\mu_2\sigma_2\rho[\phi(c_a) - \phi(c_b)] \tag{3.3}$$

$$+ \sigma^2\rho^2[c_a\phi(c_a) - c_b\phi(c_b)] \,,$$

$$\text{(iii)} \quad \iint_A x_1x_2 f(x_1,x_2)dx_1dx_2$$

$$= [\mu_2\sigma_1 + \mu_1\sigma_2\rho][\phi(c_a) - \phi(c_b)] + [\mu_1\mu_2 + \rho\sigma_1\sigma_2][\Phi(c_b) - \Phi(c_a)]$$
$$+ \sigma_1\sigma_2\rho[c_a\phi(c_a) - c_b\phi(c_b)] ,$$

$$\text{(iv)} \quad \iint_A x_1 f(x_1,x_2)dx_1dx_2 = \sigma_1[\phi(c_a) - \phi(c_b)] + \mu_1[\Phi(c_b) - \Phi(c_a)] ,$$

$$\text{(v)} \quad \iint_A x_2 f(x_1,x_2)dx_1dx_2 = \sigma_2\rho[\phi(c_a) - \phi(c_b)] + \mu_2[\Phi(c_b) - \Phi(c_a)] ,$$

where $c_a = -\mu_1/\sigma_1$ and $c_b = c - \mu_1/\sigma_1$.

Proof: (i) and (iv) follows from (3.2). To prove (ii), consider $X_2|x_1 \sim N[\mu_2 + (\sigma_2/\sigma_1)\rho(x_1-\mu_1), \sigma^2(1-\rho^2)]$ and $X_1 \sim N(\mu_1,\sigma_1^2)$. Let $f(x_2|x_1)$ be the p.d.f. of $X_2|x_1$ and $f(x_1)$ be the p.d.f. of X_1. Then

$$\text{(3.4)} \quad \iint_A x_2^2 f(x_1,x_2)dx_1dx_2 = \iint_A x_2^2 f(x_2|x_1)f(x_1)dx_2dx_1$$
$$= \int_0^c \{\sigma_2^2(1-\rho^2) + [\mu_2 + (\sigma_2/\sigma_1)\rho(x_1-\mu_1)]^2\}f(x_1)dx_1 .$$

Apply (3.2) to (3.4), we get (ii). By similar arguments, we get (iii) and (v). Q.E.D.

In Section 3.2, we have shown that if (p,τ) lies inside the ellipse (see Figures 2.1 and 2.2) then the always-pooled estimator is better than the never-pooled estimator. We use two parallel lines $\tau = \beta p$ and $\tau = \beta(p-1)$ as the linearly approximated boundary of the ellipse (see Figure 3.1). Then roughly speaking, the always-pooled estimator is better than the never-pooled estimator if (p,τ) lies between these two parallel lines. Hence we define the sometimes-pooled estimator based on the linearly approximated mean square error as follows:

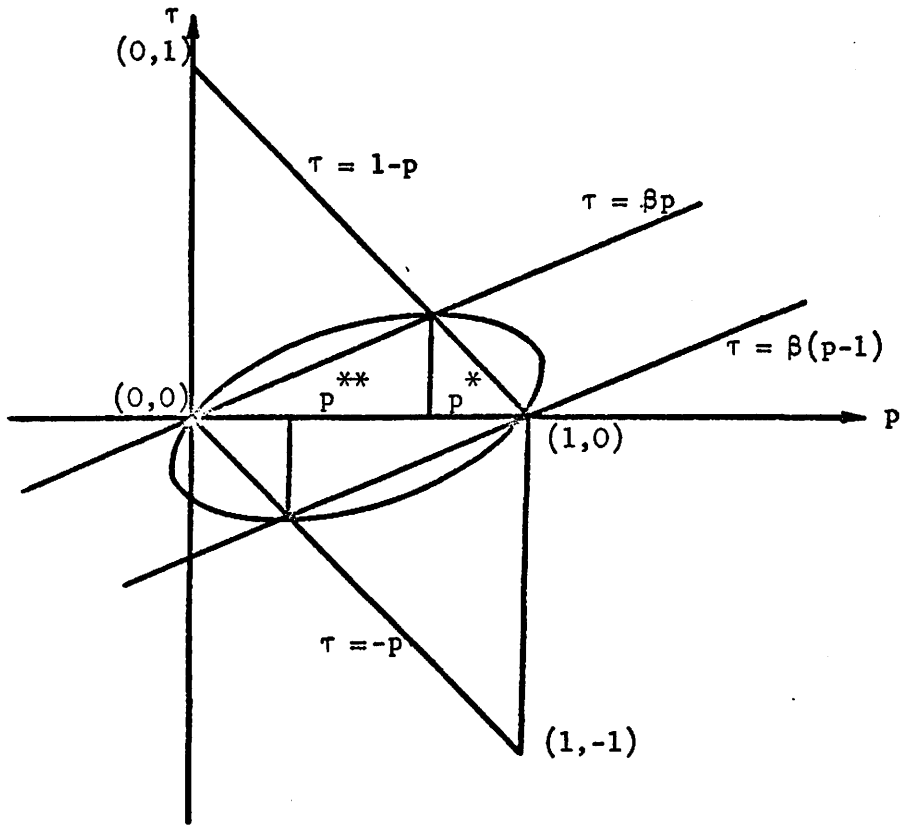


Figure 3.1 ($\beta = (n_2 + 2n_1)/(n_1 n_2)$)

$$(3.5) \qquad \hat{p}_s = \begin{cases} \hat{p}_A, & \text{if } \beta(\hat{p}-1) \le \hat{\tau} \le \beta\hat{p} \\ \hat{p}_N, & \text{otherwise} , \end{cases}$$

where

$$(3.6) \qquad \hat{p} = X_1/n_1, \ \hat{\tau} = X_2/n_2 - X_1/n_1, \text{ and } \beta = (n_2 + 2n_1)/(n_1 n_2).$$

Let

$$(3.7) \qquad \xi_1 = X_1/n_1 \text{ and } \xi_2 = X_2/n_2 .$$

Assume that n_1 and n_2 are sufficiently large such that approximately

$$(3.8) \qquad \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \sim N\left[\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right],$$

where $\mu_1 = p$, $\mu_2 = p+\tau$, $\sigma_1^2 = p(1-p)/n_1$ and $\sigma_2^2 = (p+\tau)(1-p-\tau)/n_2$. Then (3.5) can be rewritten as follows:

$$(3.9)\qquad \hat{p}_s = \begin{cases} (n_1\xi_1 + n_2\xi_2)/(n_1+n_2), & \text{if } (\xi_1,\xi_2) \in A \\ \xi_1 , & \text{if } (\xi_1,\xi_2) \in A^c , \end{cases}$$

where

$$(3.10)\qquad A = \{(\xi_1,\xi_2) \mid \xi_2/(1+\beta) \le \xi_1 \le (\xi_2+\beta)/(1+\beta)\}.$$

It follows that

$$(3.11)\qquad E(\hat{p}_s - p)^2 = \frac{p(1-p)}{n_1} + \iint_A [(\frac{n_1\xi_1+n_2\xi_2}{n_1+n_2})^2 - \xi_1^2 - 2\mu_1(\frac{n_1\xi_1+n_2\xi_2}{n_1+n_2}) + 2\mu_1\xi_1]f(\xi_1,\xi_2)d\xi_1 d\xi_2,$$

where $f(\xi_1,\xi_2)$ is the p.d.f. of (ξ_1,ξ_2). Let

$$(3.12)\qquad \eta_1 = \xi_1 - \xi_2/(1+\beta), \quad \eta_2 = \xi_2 .$$

From (3.8) it follows that approximately

$$(3.13)\qquad \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \sim N\left[\begin{bmatrix} \mu_1^* \\ \mu_2^* \end{bmatrix}, \begin{bmatrix} \sigma_1^{*2} & \sigma_{12}^* \\ \sigma_{12}^* & \sigma_2^{*2} \end{bmatrix}\right],$$

where $\mu_1^* = \mu_1 - \mu_2/(1+\beta)$, $\mu_2^* = \mu_2$, $\sigma_1^{*2} = \sigma_1^2 + \sigma_2^2/(1+\beta)^2$, $\sigma_{12}^* = -\sigma_2^2/(1+\beta)$ and $\sigma_2^{*2} = \sigma_2^2$. Replacing (ξ_1,ξ_2) by (η_1,η_2), (3.11) is equivalent to

$$(3.14)\qquad E(\hat{p}_s - p)^2 = \frac{P(1-p)}{n_1} + \iint_{A^*} [a\eta_1^2 + b\eta_2^2 + c\eta_1\eta_2 + d\eta_1 + e\eta_2]f^*(\eta_1,\eta_2)d\eta_1 d\eta_2 ,$$

where f^* is the p.d.f. of (η_1,η_2), $A^* = \{(\eta_1,\eta_2) \mid 0 \le \eta_1 \le \beta/(1+\beta)\}$, $a = -(2n_1n_2 + n_2^2)/(n_1+n_2)^2$, $b = [(2n_1n_2 + 2n_2^2)\beta + n_2^2\beta^2]/[(n_1+n_2)^2(1+\beta)^2]$, $c = (2n_1n_2\beta - 2n_1n_2 - 2n_2^2)/[(n_1+n_2)^2(1+\beta)]$, $d = 2\mu_1 n_2/(n_1+n_2)$ and

$e = -2\mu_1 n_1 \beta/[(n_1+n_2)(1+\beta)]$. Applying (3.3) and (3.13) to (3.14), we find that

$$
\begin{aligned}
& E(\hat{p}_s - p)^2 \cong \\
& [a\mu_1^{*2} + b\mu_2^{*2} + c\mu_1^*\mu_2^* + d\mu_1^* + e\mu_2^* + a\sigma_1^{*2} + b\sigma_2^{*2} + c\sigma_{12}^*] \times \\
(3.15)\quad & [\Phi(c_b^*) - \Phi(c_a^*)] \\
& + [2a\sigma_1^*\mu_1^* + 2b\mu_2^*\sigma_2^*\rho^* + c\mu_2^*\sigma_1^* + c\sigma_2^*\rho^*\mu_1^* + d\sigma_1^* + e\sigma_2^*\rho^*] \times \\
& [\phi(c_a^*) - \phi(c_b^*)] \\
& + [a\sigma_1^{*2} + b\sigma_2^{*2}\rho^{*2} + c\sigma_{12}^*] \times [c_a^*\phi(c_a^*) - c_b^*\phi(c_b^*)] ,
\end{aligned}
$$

where $\rho^* = \sigma_{12}^*/(\sigma_1^*\sigma_2^*)$, $c_a^* = -\mu_1^*/\sigma_1^*$ and $c_b = [\beta/(1+\beta) - \mu_1^*]/\sigma_1^*$.

Next, we consider the sometimes-pooled estimator in the Kale-Bancroft (1967) sense. Let

$$(3.16)\qquad \xi_1' = \sin^{-1}\sqrt{\xi_1} \quad \text{and} \quad \xi_2' = \sin^{-1}\sqrt{\xi_2} .$$

Assume that n_1 and n_2 are sufficiently large such that approximately

$$(3.17)\qquad \begin{bmatrix} \xi_1' \\ \xi_2' \end{bmatrix} \sim N\left[\begin{pmatrix} \sin^{-1}\sqrt{p} \\ \sin^{-1}\sqrt{p+\tau} \end{pmatrix}, \begin{pmatrix} 1/(4n_1) & 0 \\ 0 & 1/(4n_2) \end{pmatrix}\right].$$

It follows that approximately

$$(3.18)\qquad \xi_2' - \xi_1' \sim N[\sin^{-1}\sqrt{p+\tau} - \sin^{-1}\sqrt{p},\ (n_1+n_2)/(4n_1n_2)] .$$

We transform the problem of estimating p to the problem of estimating $\sin^{-1}\sqrt{p}$. We define the sometimes-pooled estimator based on the preliminary test of $H_0: \sin^{-1}\sqrt{p+\tau} = \sin^{-1}\sqrt{p}$ with significance level = α as follows:

$$(3.19)\qquad \hat{\mu}_{1s\alpha} = \begin{cases} (n_1\xi_1' + n_2\xi_2')/(n_1+n_2), & \text{if } (\xi_1',\xi_2') \in A_\alpha \\ \xi_1', & \text{otherwise,} \end{cases}$$

where

$$(3.20)\qquad A_\alpha = \{(\xi_1',\xi_2') \mid |(\xi_2'-\xi_1')/\sqrt{(n_1+n_2)/(4n_1n_2)}| \le c_\alpha\},$$

where c_α is the solution of $1 - \Phi(c_\alpha) = \alpha/2$. Then, it follows that

$$(3.21)\qquad \begin{aligned} &E(\hat{\mu}_{1s\alpha} - \sin^{-1}\sqrt{p})^2 \\ &= \frac{1}{4n_1} + \iint_{A_\alpha}\Big[\Big(\frac{n_1\xi_1' + n_2\xi_2'}{n_1+n_2}\Big)^2 - \xi_1'^2 - 2\sin^{-1}\sqrt{p}\,\Big(\frac{n_1\xi_1' + n_2\xi_2'}{n_1+n_2}\Big) \\ &\qquad + 2\sin^{-1}\sqrt{p}\,\xi_1'\Big] f(\xi_1',\xi_2')\,d\xi_1'\,d\xi_2', \end{aligned}$$

where $f(\xi_1',\xi_2')$ is the p.d.f. of (ξ_1',ξ_2'). Let

$$(3.22)\qquad \eta_1' = \xi_2' - \xi_1' + c_\alpha\sqrt{(n_1+n_2)/(4n_1n_2)} \quad \text{and} \quad \eta_2' = \xi_2'.$$

From (3.17) it follows that approximately

$$(3.23)\qquad \begin{bmatrix}\eta_1' \\ \eta_2'\end{bmatrix} \sim N\left[\begin{pmatrix}\mu_1' \\ \mu_2'\end{pmatrix}, \begin{pmatrix}\sigma_1'^2 & \sigma_{12}' \\ \sigma_{12}' & \sigma_2'^2\end{pmatrix}\right],$$

where $\mu_1' = \sin^{-1}\sqrt{p+\tau} - \sin^{-1}\sqrt{p} + c_\alpha\sqrt{(n_1+n_2)/(4n_1n_2)}$, $\mu_2' = \sin^{-1}\sqrt{p+\tau}$, $\sigma_1'^2 = (n_1+n_2)/(4n_1n_2)$, $\sigma_{12}' = \sigma_2'^2 = 1/(4n_2)$. Replacing (ξ_1',ξ_2') by (η_1',η_2'), (3.21) is equivalent to

$$(3.24)\qquad \begin{aligned} &E(\hat{\mu}_{1s\alpha} - \sin^{-1}\sqrt{p})^2 \\ &= \frac{1}{4n_1} + \iint_{A^*}[a'\eta_1'^2 + k'\eta_1'\eta_2' + d'\eta_1' + e'\eta_2' + f']f^*(\eta_1',\eta_2')\,d\eta_1'\,d\eta_2', \end{aligned}$$

where $A^* = \{(\eta_1', \eta_2') \mid 0 \le \eta_1' \le c'\}$, $c' = \sqrt{(n_1+n_2)/(n_1 n_2)}\, c_\alpha$, $f^*(\eta_1', \eta_2')$ is the p.d.f. of (η_1', η_2'), $a' = -(2n_1 n_2 + n_2^2)/(n_1+n_2)^2$, $k' = 2n_2/(n_1+n_2)$, $d' = 1/(n_1+n_2)^2[2n_1 n_2(c' - \sin^{-1}\sqrt{p}) + n_2^2 c' - 2n_2^2 \sin^{-1}\sqrt{p}]$, $e' = -n_2 c'/(n_1+n_2)$, and $f' = 1/[4(n_1+n_2)^2][(4n_1 n_2 + 4n_2^2)\sin^{-1}\sqrt{p}\, c' - (2n_1 n_2 + n_2^2)c'^2]$. Applying (3.3) and (3.23) to (3.24), we find that

$$
\begin{aligned}
&E(\hat{\mu}_{1s\alpha} - \sin^{-1}\sqrt{p})^2 \cong \\
&\quad [a'\mu_1'^2 + k'\mu_1'\mu_2' + d'\mu_1' + e'\mu_2' + a'\sigma_1'^2 + k'\sigma_{12}' + f'] \times \\
&\quad [\Phi(c_b') - \Phi(c_a')] \\
&\quad + [2a'\sigma_1'\mu_1' + k'\mu_2'\sigma_1' + k'\sigma_2'\rho'\mu_1' + d'\sigma_1' + e'\sigma_2'\rho'] \times \\
&\quad [\phi(c_a') - \phi(c_b')] \\
&\quad + [a'\sigma_1'^2 + k'\sigma_{12}'] \times [c_a'\phi(c_a') - c_b'\phi(c_b')] ,
\end{aligned}
\tag{3.25}
$$

where $\rho' = \sigma_{12}'/[\sigma_1'\sigma_2']$, $c_a' = -\mu_1'/\sigma_1'$ and $c_b' = [c' - \mu_1']/\sigma_1'$.

Before we compare the performance of the sometimes-pooled estimator based on a linearly approximated mean square error (3.9) to that of the sometimes-pooled estimator in the Kale - Bancroft sense (3.19), we must note the following fact.

Lemma 3.4. For all $0 < \alpha < 1$, A_α in (3.20) $\neq A$ in (3.10).

Proof: It is easy to see that A_α is symmetric in (ξ_1, ξ_2) for all $0 < \alpha < 1$. That is, $(\xi_1, \xi_2) \in A_\alpha \Leftrightarrow (\xi_2, \xi_1) \in A_\alpha$. But A is not symmetric in (ξ_1, ξ_2). The result follows. Q.E.D.

Remark: Lemma 3.4 tells us that these two sometimes-pooled estimators are completely different in the sense that they never have identical pooling regions.

Let

(3.26) $e = E(\hat{p}_N - p)^2 / E(\hat{p}_s - p)^2$,

where $\hat{p}_N$ is indicated in (2.3) and $\hat{p}_s$ is indicated in (3.9).

Let

(3.27) $e_\alpha = E(\hat{\mu}_{1N} - \sin^{-1}\sqrt{p})^2 / E(\hat{\mu}_{1s\alpha} - \sin^{-1}\sqrt{p})^2$,

where $\hat{\mu}_{1N} = \sin^{-1}\sqrt{x_1/n_1}$ and $\hat{\mu}_{1s\alpha}$ is indicated in (3.19). We will compare the values of e and e_α, for $\alpha = 0.01, 0.05, 0.10, 0.25$ and 0.50. We use normal approximation to approximate e and e_α's. We fix $n_1 = 25$ and $n_2 = 30$. The computations of $E(\hat{p}_N - p)^2$ and $E(\hat{\mu}_{1N} - \sin^{-1}\sqrt{p})^2$ are straightforward. We use (3.15) and (3.25) to compute the approximated values of $E(\hat{p}_s - p)^2$ and $E(\hat{\mu}_{1s\alpha} - \sin^{-1}\sqrt{p})^2$. Table 3.1 gives us approximated values of e and e_α when we fix $p = 0.05$. Table 3.2 gives us values when we fix $p = 0.50$. Table 3.1 ($p = 0.05$) shows that $e_{0.01}$ is better than e when τ is between 0 and 0.10. But when τ is greater than 0.15, e is better than $e_{0.01}$. Also $e_{0.05}$ is better than e when τ is between 0 and 0.05. But e is better than $e_{0.05}$ when τ is greater than 0.10. Because of symmetric property, Table 3.2 ($p = 0.50$) gives us identical values for both τ and $-\tau$. Table 3.2 shows that both $e_{0.01}$ and $e_{0.05}$ are better than e when τ is between -0.10 and 0.10 but beyond that e performs better. For large values of τ, e performs better than e_α. Because of symmetric property, Table 3.1 will give us identical values if $p = 0.05$ is replaced by $p = 0.95$ and τ is replaced by $-\tau$.

Table 3.1 ($p = 0.05$, $n_1 = 25$, $n_2 = 30$)

	e	$e_{0.01}$	$e_{0.05}$	$e_{0.10}$	$e_{0.25}$	$e_{0.50}$
$\tau = 0.00$	1.356	1.847	1.648	1.441	1.178	1.040
$\tau = 0.05$	0.983	1.546	1.110	1.060	1.014	1.001
$\tau = 0.10$	0.858	0.989	0.727	0.768	0.869	0.962
$\tau = 0.15$	0.856	0.748	0.586	0.668	0.828	0.955
$\tau = 0.20$	0.900	0.656	0.556	0.665	0.849	0.966
$\tau = 0.25$	0.947	0.673	0.588	0.716	0.894	0.980
$\tau = 0.30$	0.979	0.674	0.660	0.793	0.938	0.990
$\tau = 0.35$	0.997	0.733	0.755	0.871	0.969	0.996
$\tau = 0.40$	0.999	0.809	0.849	0.932	0.987	0.998
$\tau = 0.45$	0.999	0.882	0.921	0.969	0.995	0.999
$\tau = 0.50$	1	0.938	0.965	0.988	0.998	0.999
$\tau = 0.55$	1	0.972	0.987	0.996	0.999	1
$\tau = 0.60$	1	0.990	0.996	0.999	0.999	1
$\tau = 0.65$	1	0.997	0.999	0.999	1	1
$\tau = 0.70$	1	0.999	0.999	1	1	1
$\tau = 0.75$	1	0.999	1	1	1	1
$\tau = 0.80$	1	0.999	1	1	1	1
$\tau = 0.85$	1	1	1	1	1	1
$\tau = 0.90$	1	1	1	1	1	1

Table 3.2 ($p = 0.50$, $n_1 = 25$, $n_2 = 30$)

	e	$e_{0.01}$	$e_{0.05}$	$e_{0.10}$	$e_{0.25}$	$e_{0.50}$
$\tau = -0.45$	0.999	0.703	0.921	0.969	0.995	0.999
$\tau = -0.40$	0.996	0.493	0.764	0.877	0.971	0.996
$\tau = -0.35$	0.990	0.407	0.629	0.762	0.922	0.987
$\tau = -0.30$	0.981	0.398	0.563	0.682	0.866	0.972
$\tau = -0.25$	0.973	0.451	0.565	0.659	0.831	0.958
$\tau = -0.20$	0.968	0.574	0.638	0.702	0.838	0.956
$\tau = -0.15$	0.973	0.803	0.800	0.825	0.897	0.970
$\tau = -0.10$	1.063	1.189	1.081	1.038	1.003	1.001
$\tau = -0.05$	1.149	1.704	1.446	1.302	1.122	1.027
$\tau = 0.00$	1.188	1.847	1.648	1.441	1.178	1.040
$\tau = 0.05$	1.149	1.704	1.446	1.302	1.122	1.027
$\tau = 0.10$	1.063	1.189	1.081	1.038	1.003	1.001
$\tau = 0.15$	0.973	0.803	0.800	0.825	0.897	0.970
$\tau = 0.20$	0.968	0.574	0.638	0.702	0.838	0.956
$\tau = 0.25$	0.973	0.451	0.565	0.659	0.831	0.958
$\tau = 0.30$	0.981	0.398	0.563	0.682	0.866	0.972
$\tau = 0.35$	0.990	0.407	0.629	0.762	0.922	0.987
$\tau = 0.40$	0.996	0.493	0.764	0.877	0.971	0.996
$\tau = 0.45$	0.999	0.703	0.921	0.969	0.995	0.999

3.4. Two normal populations and two Poisson populations.

Suppose we have two normal populations, say $N_i(\mu_i,\sigma^2)$, $i = 1,2$, μ_i and σ^2 being unknown. Suppose we have the following past data available: $(Y_{ij}, j = 1,2,\ldots,n_i)$ from $N_i(\mu_i,\sigma^2)$, $i = 1,2$. Y_{ij}'s are independent. Let $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$, $i = 1,2$. We want to estimate μ_1. We have two estimators for μ_1 as follows:

(i) Never-pooled estimator $\hat{\mu}_{1N} = \bar{Y}_1$,

(ii) Always-pooled estimator $\hat{\mu}_{1A} = (n_1\bar{Y}_1 + n_2\bar{Y}_2)/(n_1 + n_2)$.

It is easy to see that

$$(4.1) \quad E(\hat{\mu}_{1A} - \mu_1)^2 < E(\hat{\mu}_{1N} - \mu_1)^2 \Leftrightarrow (\mu_1-\mu_2)^2 < (1/n_1 + 1/n_2)\sigma^2 .$$

From (4.1), we can define the sometimes-pooled estimator based on the estimated mean square error as follows:

$$(4.2) \quad \hat{\mu}_{1s} = \begin{cases} \hat{\mu}_{1A}, & \text{if } (\bar{Y}_1 - \bar{Y}_2)^2 < (1/n_1 + 1/n_2)S_p^2 \\ \hat{\mu}_{1N}, & \text{otherwise} , \end{cases}$$

where $S_p^2 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2/(n_1 + n_2 - 2)$. Also we can define the sometimes-pooled estimator based on the preliminary test of $H_0:\mu_1 = \mu_2$ with significance level = α as follows:

$$(4.3) \quad \hat{\mu}_{1s\alpha} = \begin{cases} \hat{\mu}_{1A}, & \text{if } |t| < t_\alpha \\ \hat{\mu}_{1N}, & \text{otherwise} , \end{cases}$$

where $t = (\bar{Y}_1 - \bar{Y}_2)/(S_p\sqrt{1/n_1 + 1/n_2})$ and t_α is $(1-\alpha/2)$th quantile of t distribution with $n_1 + n_2 - 2$ degrees of freedom. Then we have the following result.

Lemma 4.1. There exists an α, $0 < \alpha < 1$ such that $\hat{\mu}_{1s\alpha} = \mu_{1s}$.

Proof: (4.2) is equivalent to the following:

$$(4.4)\quad \hat{\mu}_{1s\alpha} = \begin{cases} \hat{\mu}_{1A}, & \text{if } (\bar{Y}_1 - \bar{Y}_2)^2 < (1/n_1 + 1/n_2)S_p^2\, t_\alpha^2 \\ \hat{\mu}_{1N}, & \text{otherwise .} \end{cases}$$

If we choose α such that $t_\alpha = 1$, then $\hat{\mu}_{1s\alpha} = \hat{\mu}_{1s}$ by (4.2) and (4.4). Q.E.D.

Remark: Lemma 4.1 tells us that the sometimes-pooled estimator based on the estimated mean square error is just a special case of the sometimes-pooled estimator based on the preliminary tests of significance. We have seen in Lemma 3.4 that this property does not hold in the case of two binomial populations. As we have mentioned before, the performance of $\hat{\mu}_{1s\alpha}$ has been studied by Han and Bancroft (1968).

Suppose we have two Poisson populations, say $P_i(\lambda_i)$, $i = 1,2$. Suppose we have the following past data available: $(Y_{ij}, j = 1,\ldots,n_i)$ from $P_i(\lambda_i)$, $i = 1,2$. Y_{ij}'s are independent. Let $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$, $i = 1,2$. We want to estimate λ_1. We have two estimators for λ_1 as follows:

(i) Never-pooled estimator $\hat{\lambda}_{1N} = \bar{Y}_1$,

(ii) Always-pooled estimator $\hat{\lambda}_{1A} = (n_1\bar{Y}_1 + n_2\bar{Y}_2)/(n_1 + n_2)$.

Let

$$(4.5)\quad D(\lambda_1,\tau) = E(\hat{\lambda}_{1A} - \lambda_1)^2 - E(\hat{\lambda}_{1N} - \lambda_1)^2 = a\tau^2 + b\tau + c\lambda_1 ,$$

where $\tau = \lambda_2 - \lambda_1$, $a = n_2^2/(n_1 + n_2)^2$, $b = n_2/(n_1 + n_2)^2$ and $c = -n_2/[n_1(n_1 + n_2)]$. Let

(4.6) $$D^*(\lambda_1,\tau) = \{(\lambda_1,\tau) | D(\lambda_1,\tau) = 0\}.$$

Lemma 4.2.

(i) $D^*(\lambda_1,\tau)$ is a parabola.

(4.7) (ii) For fixed τ, $D(\lambda_1,\tau) \downarrow -\infty$ as $\lambda_1 \uparrow \infty$.

(iii) For fixed λ_1, $D(\lambda_1,\tau) \uparrow \infty$ as $\tau \uparrow \infty$.

Proof: (4.7) follows from (4.5) and (4.6). Q.E.D.

As in the normal and the binomial cases, we can define the sometimes-pooled estimator based on the estimated mean square error as follows:

$$\hat{\lambda}_{1s} = \begin{cases} \hat{\lambda}_{1A} & \text{if } (\bar{Y}_1,\bar{Y}_2) \in A \\ \hat{\lambda}_{1N} & \text{otherwise} , \end{cases}$$

where

(4.8) $$A = \{(\bar{Y}_1,\bar{Y}_2) | D(\hat{\lambda}_1,\hat{\tau}) < 0\} ,$$

where $\hat{\lambda}_1 = \bar{Y}_1$ and $\hat{\tau} = \bar{Y}_2 - \bar{Y}_1$.

As we have mentioned in Chapter 2, Kale and Bancroft (1967) suggest a sometimes-pooled estimator based on the preliminary tests of significance. They use square root transformation and normal approximation. Let

(4.9) $$Y_{ij}^{(T)} = \sqrt{Y_{ij}} \quad j = 1,2,\ldots,n_i, \quad i = 1,2 .$$

Suppose that λ_1 and λ_2 are sufficiently large such that approximately

(4.10) $$Y_{ij}^{(T)} \sim N(\mu_i, \frac{1}{4}) ,$$

where $\mu_i = \sqrt{\lambda_i}$, $i = 1,2$. We transform the problem of estimating λ_1 to the problem of estimating μ_1. Let $\bar{Y}_i^{(T)} = \sum_{j=1}^{n_i} Y_{ij}^{(T)}/n_i$, $i = 1,2$. Then approximately

$$(4.11) \qquad \bar{Y}_1^{(T)} - \bar{Y}_2^{(T)} \sim N[\mu_1 - \mu_2, (n_2 + n_1)/4n_1n_2] .$$

We can define the sometimes-pooled estimator $\hat{\mu}_{1s\alpha}$ based on the preliminary tests of $H_0: \mu_1 = \mu_2$ with significance level = α as follows:

$$(4.12) \qquad \hat{\mu}_{1s\alpha} = \begin{cases} (n_1\bar{Y}_1^{(T)} + n_2\bar{Y}_2^{(T)})/(n_1 + n_2) & \text{if } (\bar{Y}_1^{(T)}, \bar{Y}_2^{(T)}) \in A_\alpha \\ \bar{Y}_1^{(T)} & \text{otherwise} , \end{cases}$$

where

$$(4.13) \qquad A_\alpha = \{(\bar{Y}_1^{(T)}, \bar{Y}_2^{(T)}) \mid |(\bar{Y}_1^{(T)} - \bar{Y}_2^{(T)})/\sqrt{(n_1 + n_2)/(4n_1n_2)}| \le c_\alpha\} ,$$

where c_α is the solution of $1 - \Phi(c_\alpha) = \alpha/2$. Then we have the following result.

Lemma 4.3. For all $0 < \alpha < 1$, A_α in (4.13) $\neq A$ in (4.8).

Proof: (4.13) is equivalent to the following:

$$(4.14) \quad A_\alpha = \{(Y_{ij}^{(T)}, j=1,2,\ldots,n_i, i=1,2) \mid (\bar{Y}_1^{(T)} - \bar{Y}_2^{(T)})/\sqrt{(n_1+n_2)/(4n_1n_2)}| \le c_\alpha\} ,$$

where

$$(4.15) \qquad \bar{Y}_i^{(T)} \text{ is a linear function of } Y_{ij}^{(T)}.$$

It is easy to see that (4.8) is equivalent to the following:

$$(4.16) \qquad A = \{(Y_{ij}^{(T)}, j = 1,2,\ldots,n_i, i = 1,2) \mid a(\bar{Y}_2 - \bar{Y}_1)^2 + b(\bar{Y}_2 - \bar{Y}_1) + c\bar{Y}_1 < 0\} ,$$

where

$$(4.17) \qquad \bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}^{(T)2}/n_i \text{ is not a linear function of } Y_{ij}^{(T)} .$$

The result follows from (4.15) and (4.17). Q.E.D.

Remark: Lemma 4.3 tells us that the sometimes-pooled estimator based on the estimated mean square error and the sometimes-pooled estimator in the Kale - Bancroft sense are different in the sense that they never have identical pooling regions. This phenomenon is analogous to that of two binomial populations case.

CHAPTER 4

A DECISION THEORETIC APPROACH TO THE PROBLEM OF POOLING DATA (TWO POPULATIONS CASE)

4.1. Introduction.

Suppose that we have past data available from two populations whose parameters are θ_1 and θ_2. We have two estimators for θ_1 as follows: (i) the never-pooled estimator $\hat{\theta}_{1N}$, which depends on the observations from the first population only, (ii) the always-pooled estimator $\hat{\theta}_{1A}$, which depends on the observations from both populations. It is well known that $E(\hat{\theta}_{1A} - \theta_1)^2$ is sometimes less than $E(\hat{\theta}_{1N} - \theta_1)^2$. We also define a sometimes-pooled estimator which is equal to the always-pooled estimator according to a certain pooling rule.

It seems that all discussions in the literature on pooling of data are confined to the problem of finding the best estimator of θ_1. In Chapters 2 and 3 we have used the mean square error as criterion to evaluate the performances of the various estimators for θ_1, the always-pooled, the never-pooled and the sometimes-pooled. But so far all discussions have stopped at this point and haven't gone beyond that. We never ask the following question: why do we want to find an estimator for θ_1? The following discussions attempt to answer this question. We try to discuss the problem of pooling data beyond the framework of estimation.

Simply, the reason for finding an estimator for θ_1 is that we want to make a "decision" (or a "prediction"). Imagine that we have a given population of animals which can be classified into two

categories, I and II. Category I is composed of animals which respond to a particular medical treatment. Category II is composed of animals which do not respond to the treatment. Let p be the probability that a certain animal belongs to category I. We consider the case where an always-pooled and a never-pooled estimators for p are available. Suppose that we have a new animal from the given population. We want to make a treatment decision: apply treatment to it or not. If we know which category the animal belongs to, then the decision is easy. However, in practice we may not know which category it belongs to. Consequently, our decision rule maight reasonably depend on some estimated $\hat{p}$ of p. In particular, we consider the always-pool, the never-pool, and the sometimes-pool decision rules according to the corresponding estimators we use. We find out that in this framework of discussion the pooling of data is irrelevant in a certain sense. In other words, the pooling of data may help us find a better estimator, but when we use it for decision making it does not make much difference.

In Section 4.2, we discuss the case of two binomial populations and two treatments. In Section 4.3, we discuss the arcsine square root transformation in the same case. In Section 4.4, we discuss the case of two binomial populations and s treatments. In Section 4.5, we discuss the case of two r-variate multinomial populations and two treatments. In Section 4.6, we duscuss the problem of prediction in the case of two normal populations. In Section 4.7, we discuss the implications of the main result which seem a bit subtle.

4.2. 2 x 2 case.

Suppose we have two binomial populations, say π_1 and π_2. Also we have two kinds of individuals in each population, say α (responding to a particular treatment) and β (not responding to the treatment). Let

(2.1) $p_i = P(\alpha|\pi_i)$, $q_i = 1-p_i$, $0 < p_i < 1$, $i = 1,2$.

Suppose we have the following past data available:

(2.2) $X_i\alpha$ individuals out of total n_i observations from π_i, $i = 1,2$; X_1 and X_2 are independent.

Suppose Y_1 is a new individual from π_1. Y_1 is independent of X_i. We don't know if Y_1 is α or β. We want to make a treatment decision: apply treatment to Y_1 or not. Suppose we have the following 2 x 2 loss table.

	treatment	no treatment
Y_1 being α	a	b
Y_1 being β	c	d

Table 2.1

Treatment is better for α individuals, since they respond to the treatment. Non-treatment is better for β individuals, since they do not respond to the treatment. Hence we can assume that a, b, c and d are arbitrary real numbers satisfying the following inequalities:

(2.3) $a < b$ and $d < c$.

It follows that

$$(2.4)\quad \begin{cases} \text{Expected loss when applying treatment} = p_1 a + q_1 c \\ \text{Expected loss when applying no treatment} = p_1 b + q_1 d \ . \end{cases}$$

Hence we can define a treatment decision rule based on expected loss as follows:

$$(2.5)\quad \begin{cases} \text{Treat if } \ \tilde{p}_1 a + \tilde{q}_1 c < \tilde{p}_1 b + \tilde{q}_1 d \\ \text{Don't treat if } \ \tilde{p}_1 a + \tilde{q}_1 c \geq \tilde{p}_1 b + \tilde{q}_1 d \ , \end{cases}$$

where $\tilde{p}_1$ is an estimator of p_1, $\tilde{q}_1 = 1-\tilde{p}_1$, and $\tilde{p}_1$ is independent of Y_1. After some algebra, (2.5) is equivalent to

$$(2.6)\quad \begin{cases} \text{Treat if } \ \tilde{p}_1 > \delta \\ \text{Don't treat if } \ \tilde{p}_1 \leq \delta \ , \end{cases}$$

where

$$(2.7)\quad \delta = (c-d)/[(b-a) + (c-d)], \quad 0 < \delta < 1 \ .$$

Let

$$(2.8)\quad \hat{p}_i = X_i/n_i, \quad i = 1,2.$$

We define three treatment decision rules as follows:

(i) Never-pool treatment decision rule: put $\tilde{p}_1 = \hat{p}_{1N} = \hat{p}_1$.

(ii) Always-pool treatment decision rule: put

$$(2.9)\quad \tilde{p}_1 = \hat{p}_{1A} = (n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1+ n_2).$$

(iii) Sometimes-pool treatment decision rule: put $\tilde{p}_1 = \hat{p}_{1s}$, where $\hat{p}_{1s}$ depends on the pooling decision rule which will be discussed later.

Lemma 2.1. The expected loss using $\tilde{p}_1$ is

$$(2.10)\quad R(\tilde{p}_1) = c + (a-c)p_1 + \{[(b-a) + (c-d)]p_1 - (c-d)\}P[\tilde{p}_1 \leq \delta].$$

Proof: See Appendix C.

Corollary 2.1. When $p_1 = \delta$, $R(\tilde{p}_1) = c + (a-c)p_1$, the same for any $\tilde{p}_1$.

Proof: It follows immediately from Lemma 2.1.

Remark: We infer that in this case $(p_1 = \delta)$ there is nothing we can gain (or lose) by a treatment decision rule.

Assume that n_1 and n_2 are sufficiently large such that approximately

$$(2.11) \qquad \hat{p}_{1N} \sim N(\mu_1, \sigma_1^2) \quad \text{and} \quad \hat{p}_{1A} \sim N(\mu_2, \sigma_2^2) ,$$

where

$$(2.12) \qquad \mu_1 = p_1,\ \sigma_1^2 = [p_1(1-p_1)]/n_1,\ \mu_2 = (n_1p_1 + n_2p_2)/(n_1 + n_2) \text{ and}$$
$$\sigma_2^2 = [n_1p_1(1-p_1) + n_2p_2(1-p_2)]/(n_1 + n_2)^2 .$$

It follows that

$$P(\hat{p}_{1N} \le \delta) \cong \Phi[(\delta-\mu_1)/\sigma_1] \text{ and } P(\hat{p}_{1A} \le \delta) \cong \Phi[(\delta-\mu_2)/\sigma_2] .$$

In this chapter Φ always stands for the standard normal c.d.f..

We defined approximate expected losses by:

$$(2.13) \qquad \tilde{R}(\hat{p}_{1N}) = c + (a-c)p_1 + \{[(b-a)+(c-d)]p_1 - (c-d)\}\Phi[(\delta-\mu_1)/\sigma_1]$$
$$\tilde{R}(\hat{p}_{1A}) = c + (a-c)p_1 + \{[(b-a)+(c-d)]p_1 - (c-d)\}\Phi[(\delta-\mu_2)/\sigma_2] .$$

From (2.10) and (2.13), it follows that

$$(2.14) \qquad \tilde{R}(\hat{p}_{1N}) \cong R(\hat{p}_{1N}) \quad \text{and} \quad \tilde{R}(\hat{p}_{1A}) \cong R(\hat{p}_{1A}) .$$

Theorem 2.1.

$$(2.15) \quad \tilde{R}(\hat{p}_{1A}) < \tilde{R}(\hat{p}_{1N}) \Leftrightarrow (p_1, p_2) \in A_1 \cup A_2 ,$$

where $A_1 = C \cap B$, $A_2 = \bar{C} \cap \bar{B}$, $C = (\mu_1 < \delta)$, $\bar{C} = (\mu_1 > \delta)$, $B = \{\sigma_2\mu_1 - \sigma_1\mu_2 > \delta(\sigma_2 - \sigma_1)\}$ and $\bar{B} = \{\sigma_2\mu_1 - \sigma_1\mu_2 < \delta(\sigma_2 - \sigma_1)\}$.

Proof: By (2.13),

$$\tilde{R}(\hat{p}_{1A}) - \tilde{R}(\hat{p}_{1N}) = \{[(b-a) + (c-d)]\mu_1 - (c-d)\}\{\Phi[(\delta-\mu_2)/\sigma_2] - \Phi[(\delta-\mu_1)/\sigma_1]\}.$$

It follows that $\tilde{R}(\hat{p}_{1A}) < \tilde{R}(\hat{p}_{1N})$ if and only if either (i) $\mu_1 < \delta$ and $\Phi[(\delta-\mu_2)/\sigma_2] > \Phi[(\delta-\mu_1)/\sigma_1]$ or (ii) $\mu_1 > \delta$ and $\Phi[(\delta-\mu_2)/\sigma_2] < \Phi[(\delta-\mu_1)/\sigma_1]$. But $\Phi[(\delta-\mu_2)/\sigma_2] > \Phi[(\delta-\mu_1)/\sigma_1] \Leftrightarrow (\delta-\mu_2)/\sigma_2 > (\delta-\mu_1)/\sigma_1 \Leftrightarrow \sigma_2\mu_1 - \sigma_1\mu_2 > \delta(\sigma_2 - \sigma_1)$. Q.E.D.

Corollary 2.2. If $p_1 = p_2$, then $\tilde{R}(\hat{p}_{1A}) < \tilde{R}(\hat{p}_{1N})$.

Proof: It follows directly from Theorem 2.1.

Theorem 2.1 gives us the set of parameters for which $\tilde{R}(\hat{p}_{1A}) < \tilde{R}(\hat{p}_{1N})$, assuming that n_1 and n_2 are sufficiently large to justify the normal approximation. We can call this set of parameters the "pooling region". An intuitively appealing pooling decision rule is that we set $\hat{p}_{1s} = \hat{p}_{1N}$, if the estimated parameters fall in the "pooling region". Hence we have the following definition.

Definition 2.1. We define a sometimes-pool treatment decision rule based on expected loss as follows:

$$(2.16) \quad \begin{cases} \text{Treat if } \hat{p}_{1s} > \delta \\ \text{Don't treat if } \hat{p}_{1s} \leq \delta, \end{cases}$$

where the pooling decision rule is as follows:

$$(2.17)\quad \hat{p}_{1s} = \begin{cases} \hat{p}_{1A}, & \text{if } (\hat{p}_1, \hat{p}_2) \in \hat{A}_1 \cup \hat{A}_2 \\ \hat{p}_{1N}, & \text{otherwise}, \end{cases}$$

where

(2.18) $\hat{A}_i$ is the same as A_i in (2.15) except that we replace $(\mu_1,\mu_2,\sigma_1,\sigma_2)$ by $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$; while $\hat{\mu}_i$ and $\hat{\sigma}_i$ are obtained from μ_i and σ_i in (2.12) by replacing p_i by $\hat{p}_i$.

The shaded region of Figure 2.1 is the region of $\hat{A}_1 \cup \hat{A}_2$ when we fix $n_1 = 25$, $n_2 = 30$ and $\delta = 1/2$.

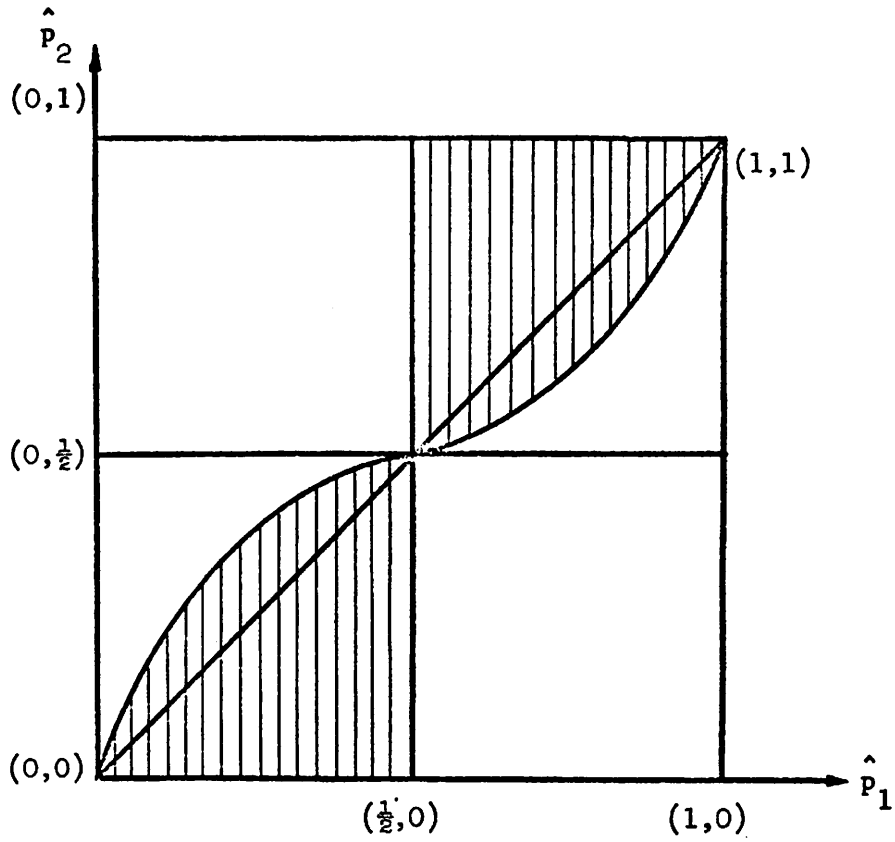


Figure 2.1

Lemma 2.2. The never-pool treatment decision rule and the always-pool treatment decision rule have different treatment decisions if and only if

$$(2.19) \qquad (\hat{p}_1, \hat{p}_2) \in \hat{B}_1 \cup \hat{B}_2 ,$$

where

$$(2.20) \qquad \hat{B}_1 = \{\hat{p}_1 \le \delta < (n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1 + n_2)\} \quad \text{and}$$
$$\hat{B}_2 = \{(n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1 + n_2) \le \delta < \hat{p}_1\} .$$

Proof: It follows directly from (2.6) and (2.9).

The shaded region of the following figure is the region of $\hat{B}_1 \cup \hat{B}_2$ when we fix $\delta = 1/2$ and $n_1 = n_2 = n$.

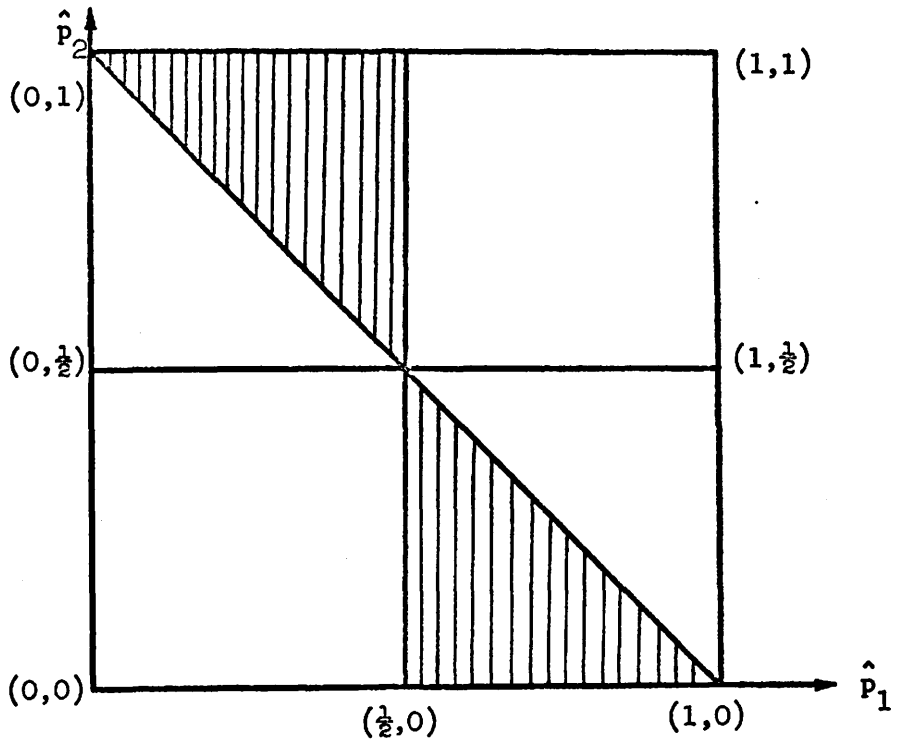


Figure 2.2

We note that the shaded regions of Figures 2.1 and 2.2 do not overlap. In other words, Figures 2.1 and 2.2 indicate that every time we use $\hat{p}_{1A}$ is when $\hat{p}_{1A}$ and $\hat{p}_{1N}$ make the same decision. Hence we may conjecture that $\hat{p}_{1s}$ in (2.17) and $\hat{p}_{1N}$ always make the same treatment decision. The following theorem shows that this conjecture is true.

Theorem 2.2. The never-pool treatment decision rule and the sometimes-pool treatment decision rule (defined in Definition 2.1) always make the same treatment decision.

Proof: Let $\hat{A} = \hat{A}_1 \cup \hat{A}_2$ and $\hat{B} = \hat{B}_1 \cup \hat{B}_2$. By (2.17) and (2.19), it is sufficient to show that $\hat{A} \subset \hat{B}^c$. This is equivalent to showing that $\hat{B} \subset \hat{A}^c$. First to show that $\hat{B}_1 \subset \hat{A}^c$. Suppose $\hat{p}_1 \leq \delta < (n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1 + n_2)$. Case (i) $\hat{p}_1 = \delta$. It is clear that $(\hat{p}_1, \hat{p}_2) \in \hat{A}^c$. Case (ii) $\hat{p}_1 < \delta < (n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1 + n_2)$. Then $\hat{\sigma}_2(\hat{\mu}_1 - \delta) < 0 < \hat{\sigma}_1(\hat{\mu}_2 - \delta)$. It follows that $\hat{\sigma}_2\hat{\mu}_1 - \hat{\sigma}_1\hat{\mu}_2 < \delta(\hat{\sigma}_2 - \hat{\sigma}_1)$. Therefore, $(\hat{p}_1, \hat{p}_2) \in \hat{A}^c$. We have shown that $\hat{B}_1 \subset \hat{A}^c$. Similarly we can show that $\hat{B}_2 \subset \hat{A}^c$. The result follows. Q.E.D.

Corollary 2.3. $R(\hat{p}_{1N}) = R(\hat{p}_{1s})$.

Proof: Since $\hat{p}_{1N}$ and $\hat{p}_{1s}$ always make the same treatment decision, their expected losses are equal. Q.E.D.

We may discuss the problem of prediction in the similar fashion. Instead of making a treatment decision on Y_1, we want to predict if Y_1 is α or β. We are subject to two kinds of error:

(i) We predict that Y_1 is α when in fact Y_1 is β.

(ii) We predict that Y_1 is β when in fact Y_1 is α.

Hence we have the following 2 x 2 loss table.

		Predicting that Y_1 is α	Predicting that Y_1 is β
Y_1 being α		0	$b > 0$
Y_1 being β		$c > 0$	0

Table 2.2

It is easy to see that the prediction problem is a special case of the treatment decision problem we have already discussed. Table 2.2 is a special case of Table 2.1. Predicting that Y_1 is α (β) corresponds to applying treatment (no treatment) to Y_1.

4.3. Arcsine square root transformation in the 2 x 2 case.

Assume that the formulation of problem and basic assumptions are the same as in Section 4.2. Let

$$(3.1) \qquad p_1^{(T)} = \sin^{-1}\sqrt{p_1} \ .$$

Then

$$(3.2) \qquad p_1 = (\sin p_1^{(T)})^2 \ .$$

It follows that (2.4) is equivalent to the following:

$$(3.3) \quad \begin{cases} \text{Expected loss when applying treatment} = [\sin p_1^{(T)}]^2 a + [\cos p_1^{(T)}]^2 c \\ \text{Expected loss when applying no treatment} = \\ \qquad\qquad [\sin p_1^{(T)}]^2 b + [\cos p_1^{(T)}]^2 d \ . \end{cases}$$

Hence we can define a treatment decision rule based on expected loss as follows:

$$(3.4) \quad \begin{cases} \text{Treat if } [\sin \tilde{p}_1^{(T)}]^2 a + [\cos \tilde{p}_1^{(T)}]^2 c < [\sin \tilde{p}_1^{(T)}]^2 b + [\cos \tilde{p}_1^{(T)}]^2 d \\ \text{Don't treat if } [\sin \tilde{p}_1^{(T)}]^2 a + [\cos \tilde{p}_1^{(T)}]^2 c \geq [\sin \tilde{p}_1^{(T)}]^2 b + [\cos \tilde{p}_1^{(T)}]^2 d, \end{cases}$$

where $\tilde{p}_1^{(T)}$ is an estimator of $p_1^{(T)}$ which is independent of Y_1.

As in the calculation of (2.6), (3.4) is equivalent to

$$(3.5)\quad \begin{cases} \text{Treat if } (\sin \tilde{p}_1^{(T)})^2 > \delta \\ \text{Don't treat if } (\sin \tilde{p}_1^{(T)})^2 \le \delta \ , \end{cases}$$

where δ is the same as in (2.7). We define three treatment decision rules as follows:

(i) Never-pool treatment decision rule: put $\tilde{p}_1^{(T)} = \hat{p}_{1N}^{(T)} = \sin^{-1}\sqrt{\hat{p}_1}$.

(ii) Always-pool treatment decision rule: put $\tilde{p}_1^{(T)} = \hat{p}_{1A}^{(T)} = \sum_{i=1}^{2} n_i \sin^{-1}\sqrt{\hat{p}_i}/(n_1+n_2)$. Note $\hat{p}_{1A}^{(T)}$ so defined has a uniformly minimum asymptotic variance among the estimators

$$(3.6)\quad \lambda \sin^{-1}\sqrt{\hat{p}_1} + (1-\lambda)\sin^{-1}\sqrt{\hat{p}_2}, \quad 0 \le \lambda \le 1.$$

(iii) Sometimes-pool treatment decision rule: put $\tilde{p}_1^{(T)} = \hat{p}_{1s}^{(T)}$, where $\hat{p}_{1s}^{(T)}$ depends on the pooling decision rule which will be discussed later.

$\hat{p}_1$ and $\hat{p}_2$ are the same as in (2.8).

Lemma 3.1. The expected loss using $\tilde{p}_1^{(T)}$ is

$$(3.7)\quad R(\tilde{p}_1^{(T)}) = c+(a-c)p_1+ \{[(b-a)+(c-d)]p_1 - (c-d)\}P[\tilde{p}_1^{(T)} \le \sin^{-1}\sqrt{\delta}] \ .$$

Proof: Similar to the proof of Lemma 2.1.

Remark: If $\tilde{p}_1^{(T)} = \sin^{-1}\sqrt{\tilde{p}_1}$, then $R(\tilde{p}_1^{(T)}) = R(\tilde{p}_1)$. However, $E(\tilde{p}_1^{(T)} - p_1^{(T)})^2$ is not the same as $E(\tilde{p}_1 - p_1)^2$ even though $\tilde{p}_1^{(T)} = \sin^{-1}\sqrt{\tilde{p}_1}$.

Theorem 3.1. Let $\tilde{p}_1$ and $\tilde{p}_1^{(T)}$ be estimators of p_1 and $p_1^{(T)}$ respectively. Then

(i) Suppose $p_1 = \delta$. Then $R(\tilde{p}_1) = R(\tilde{p}_1^{(T)}) = c + (a-c)p_1$.

(ii) Suppose $p_1 \neq \delta$. Then $R(\tilde{p}_1) = R(\tilde{p}_1^{(T)})$ for all $a,b,c,d \Rightarrow \tilde{p}_1^{(T)}$ and $\sin^{-1}\sqrt{\tilde{p}_1}$ have the same distribution.

Proof: (i) is clear. (ii) Suppose $p_1 \neq \delta$. $\tilde{p}_1^{(T)}$ and $\sin^{-1}\sqrt{\tilde{p}_1}$ have the same distribution $\Rightarrow (\sin \tilde{p}_1^{(T)})^2$ and $\tilde{p}_1$ have the same distribution $\Rightarrow P(\tilde{p}_1 \leq \delta) = P[(\sin \tilde{p}_1^{(T)})^2 \leq \delta] \Rightarrow R(\tilde{p}_1) = R(\tilde{p}_1^{(T)})$. Similarly, $R(\tilde{p}_1) = R(\tilde{p}_1^{(T)}) \Rightarrow P(\tilde{p}_1 \leq \delta) = P[(\sin \tilde{p}_1^{(T)})^2 \leq \delta]$. Since a,b,c and d are arbitrary (We only require that $a < b$ and $d < c$), δ is arbitrary. It follows that $\tilde{p}_1$ and $(\sin \tilde{p}_1^{(T)})^2$ have the same distribution. Consequently, $\sin^{-1}\sqrt{\tilde{p}_1}$ and $\tilde{p}_1^{(T)}$ have the same distribution. Q.E.D.

Corollary 3.1. Suppose $p_1 \neq \delta$. Then, in general $R(\hat{p}_{1A}) \neq R(\hat{p}_{1A}^{(T)})$.

Proof: In general, $\sin^{-1}\sqrt{\hat{p}_{1A}}$ and $\hat{p}_{1A}^{(T)}$ do not have the same distribution. By Theorem 3.1, $R(\hat{p}_{1A}) \neq R(\hat{p}_{1A}^{(T)})$. Q.E.D.

Assume that n_1 and n_2 are sufficiently large such that approximately

$$\hat{p}_{1N}^{(T)} \sim N(\mu_1, \sigma_1^2) \quad \text{and} \quad \hat{p}_{1A}^{(T)} \sim N(\mu_2, \sigma_2^2), \tag{3.8}$$

where

$$\mu_1 = \sin^{-1}\sqrt{p_1}, \quad \mu_2 = (\Sigma_{i=1}^{2} n_i \sin^{-1}\sqrt{p_i})/(n_1 + n_2), \tag{3.9}$$

$$\sigma_1^2 = 1/(4n_1) \quad \text{and} \quad \sigma_2^2 = 1/[4(n_1 + n_2)].$$

We define approximate expected losses by

$$\begin{aligned} \tilde{R}(\hat{p}_{1N}^{(T)}) &= c + (a-c)p_1 + \{[(b-a)+(c-d)]p_1 - (c-d)\}\Phi[(\sin^{-1}\sqrt{\delta} - \mu_1)/\sigma_1] \\ \tilde{R}(\hat{p}_{1A}^{(T)}) &= c + (a-c)p_1 + \{[(b-a)+(c-d)]p_1 - (c-d)\}\Phi[(\sin^{-1}\sqrt{\delta} - \mu_2)/\sigma_2]. \end{aligned} \tag{3.10}$$

From (3.7) and (3.10), it follows that

$$(3.11)\qquad \tilde{R}(\hat{p}_{1N}^{(T)}) \cong R(\hat{p}_{1N}^{(T)}) \quad \text{and} \quad \tilde{R}(\hat{p}_{1A}^{(T)}) \cong R(\hat{p}_{1A}^{(T)}).$$

Theorem 3.2.

$$(3.12)\qquad \tilde{R}(\hat{p}_{1A}^{(T)}) < \tilde{R}(\hat{p}_{1N}^{(T)}) \Leftrightarrow (p_1,p_2) \in A_1 \cup A_2 ,$$

where $A_1 = C \cap B$, $A_2=\bar{C} \cap \bar{B}$, $C = (\mu_1 < \sin^{-1}\sqrt{\delta})$, $\bar{C} = (\mu_1 > \sin^{-1}\sqrt{\delta})$, $B = \{\sigma_2\mu_1 - \sigma_1\mu_2 > \sin^{-1}\sqrt{\delta}(\sigma_2 - \sigma_1)\}$ and $\bar{B} = \{\sigma_2\mu_1 - \sigma_1\mu_2 < \sin^{-1}\sqrt{\delta}(\sigma_2-\sigma_1)\}$.

Proof: Similar to the proof of Theorem 2.1.

Corollary 3.2. If $p_1 = p_2$, then $\tilde{R}(\hat{p}_{1A}^{(T)}) < \tilde{R}(\hat{p}_{1N}^{(T)})$.

Proof: It follows directly from Theorem 3.2.

Next, we would like to define a sometimes-pool treatment decision rule based on the preliminary test of $H_0:\sin^{-1}\sqrt{p_1} = \sin^{-1}\sqrt{p_2}$ with level of significance = α.

Definition 3.1. We define a sometimes-pool treatment decision rule based on a preliminary test as follows:

$$(3.13)\qquad \begin{cases} \text{Treat if } (\sin \hat{p}_{1s\alpha}^{(T)})^2 > \delta \\ \text{Don't treat if } (\sin \hat{p}_{1s\alpha}^{(T)})^2 \le \delta , \end{cases}$$

where the pooling decision rule is as follows:

$$(3.14)\qquad \hat{p}_{1s\alpha}^{(T)} = \begin{cases} \hat{p}_{1A}^{(T)} & \text{if } (\hat{p}_1,\hat{p}_2) \in \hat{A}_\alpha \\ \hat{p}_{1N}^{(T)} & \text{otherwise} , \end{cases}$$

where

$$(3.15)\qquad \hat{A}_\alpha = \{(\hat{p}_1,\hat{p}_2) \mid |(\sin^{-1}\sqrt{\hat{p}_2} - \sin^{-1}\sqrt{\hat{p}_1})/\sqrt{(n_1+n_2)/(4n_1n_2)}| \le c_\alpha\} ,$$

where c_α is the solution of $1 - \Phi(c_\alpha) = \alpha/2$.

As in Section 4.2, we can also define a sometimes-pool treatment decision rule based on the expected loss.

Definition 3.2. We define a sometimes-pool treatment decision rule based on expected loss as follows:

$$(3.16) \qquad \begin{cases} \text{Treat if } (\sin \hat{p}_{1s}^{(T)})^2 > \delta \\ \text{Don't treat if } (\sin \hat{p}_{1s}^{(T)})^2 \le \delta, \end{cases}$$

where the pooling decision rule is as follows:

$$(3.17) \qquad \hat{p}_{1s}^{(T)} = \begin{cases} \hat{p}_{1A}^{(T)}, & \text{if } (\hat{p}_1, \hat{p}_2) \in \hat{A} \\ \hat{p}_{1N}^{(T)}, & \text{otherwise}, \end{cases}$$

where

(3.18) $\hat{A} = U_{i=1}^{2} \hat{A}_i$; $\hat{A}_i$ is the same as A_i in (3.12) except that we replace p_i by $\hat{p}_i$.

Lemma 3.2. For all $0 < \alpha < 1$, $\hat{A}_\alpha$ in (3.15) $\neq \hat{A}$ in (3.18).

Proof: Similar to the proof of Lemma 3.4 in Chapter 3.

Remark: Lemma 3.2 tells us that the pooling decision rule based on a preliminary test and the pooling decision rule based on the expected loss are completely different.

Lemma 3.3. The never-pool treatment decision rule and the always-pool treatment decision rule have different treatment decisions if and only if

$$(3.19) \qquad (\hat{p}_1, \hat{p}_2) \in \hat{B}_1 \cup \hat{B}_2,$$

where

$$B_1 = (\hat{p}_{1N}^{(T)} \le \sin^{-1}\sqrt{\delta} < \hat{p}_{1A}^{(T)}) \quad \text{and} \quad B_2 = (\hat{p}_{1A}^{(T)} \le \sin^{-1}\sqrt{\delta} < \hat{p}_{1N}^{(T)}) \ .$$

Proof: It follows directly from (3.5) and (3.6).

Theorem 3.3. The never-pool treatment decision rule and the sometimes-pool treatment decision rule based on expected loss (defined in Definition 3.2) always make the same treatment decision.

Proof: Similar to the proof of Theorem 2.2.

Corollary 3.3. $R(\hat{p}_{1N}^{(T)}) = R(\hat{p}_{1s}^{(T)})$, where $\hat{p}_{1s}^{(T)}$ is as in (3.17).

Proof: It follows directly from Theorem 3.3.

Finally, we would like to do a numerical study concerning the performance of a pooling decision rule based on the preliminary test. We define that

$$(3.20) \qquad e_\alpha = \frac{R(\hat{p}_{1N}^{(T)})}{R(\hat{p}_{1s\alpha}^{(T)})}$$

$$= \frac{c + (a-c)p_1 + \{[(b-a)+(c-d)]p_1 - (c-d)\}P(\hat{p}_{1N}^{(T)} \le \sin^{-1}\sqrt{\delta})}{c + (a-c)p_1 + \{[(b-a)+(c-d)]p_1 - (c-d)\}P(\hat{p}_{1s\alpha}^{(T)} \le \sin^{-1}\sqrt{\delta})} \ .$$

We are to use normal approximation to approximate e_α. We fix that $n_1 = 25$, $n_2 = 30$, $a = 5$, $b = 10$, $d = 5$ and $c = 10$. Then $\delta = \frac{1}{2}$. It follows that when $p_1 = \frac{1}{2}$, $e_\alpha = 1$ for all p_2 and α (as evident by Corollary 2.1 in Section 4.2). It is clear that

$$(3.21) \qquad P(\hat{p}_{1N}^{(T)} \le \sin^{-1}\sqrt{\delta}) \tilde{=} \Phi[(\sin^{-1}\sqrt{\delta} - \mu_1)/\sigma_1]$$

and

$$P(\hat{p}^{(T)}_{1s\alpha} \le \sin^{-1}\sqrt{\delta})$$

(3.22)

$$= P[\textstyle\sum_{i=1}^{2} n_i \sin^{-1}\sqrt{\hat{p}_i}/(n_1+n_2) \le \sin^{-1}\sqrt{\delta} \text{ and } |\sin^{-1}\sqrt{\hat{p}_2}-\sin^{-1}\sqrt{\hat{p}_1}| \le c]$$

$$+ P[\sin^{-1}\sqrt{\hat{p}_1} \le \sin^{-1}\sqrt{\delta} \text{ and } |\sin^{-1}\sqrt{\hat{p}_2} - \sin^{-1}\sqrt{\hat{p}_1}| > c] ,$$

where $c = \sqrt{(n_1+n_2)/(4n_1n_2)}\,c_\alpha$. It can be shown that

$$P[\textstyle\sum_{i=1}^{2} n_i \sin^{-1}\sqrt{\hat{p}_i}/(n_1+n_2) \le \sin^{-1}\sqrt{\delta} \text{ and } |\sin^{-1}\sqrt{\hat{p}_2} - \sin^{-1}\sqrt{\hat{p}_1}| \le c]$$

(3.23)

$$\tilde{=} \Phi[(\sin^{-1}\sqrt{\delta} - \mu_1^*)/\sigma_1^*]\{\Phi((2c-\mu_2^*)/\sigma_2^*) - \Phi(-\mu_2^*/\sigma_2^*)\} ,$$

where $\mu_1^* = \sum_{i=1}^{2} n_i\mu_i/(n_1+n_2)$, $\mu_2^* = \mu_1-\mu_2+c$, $\sigma_1^* = 1/(2\sqrt{n_1+n_2})$ and $\sigma_2^* = \sqrt{(n_1+n_2)/(4n_1n_2)}$. Also it can be shown that

(3.24) $$P[\sin^{-1}\sqrt{\hat{p}_1} \le \sin^{-1}\sqrt{\delta} \text{ and } |\sin^{-1}\sqrt{\hat{p}_2} - \sin^{-1}\sqrt{\hat{p}_1}| > c]$$

$$\tilde{=} \Phi[(\sin^{-1}\sqrt{\delta} - \mu_1)/\sigma_1] - \iint_R 1/[(2\pi)\sigma_1\sigma_2]e^{-(y_1-\mu_1)^2/(2\sigma_1^2)-(y_2-\mu_2)^2/(2\sigma_2^2)}dy_1dy_2,$$

where $R = (y_1 \le \sin^{-1}\sqrt{\delta}$ and $|y_2-y_1| \le c)$. We can use (3.21), (3.23) and (3.24) to compute approximated values of e_α's for $\alpha = 0.01$, 0.05, 0.10, 0.25, and 0.50. Let $p_1 = p$ and $p_2 = p+\tau$. Table 3.1 gives us values of e_α when we fix $p = 0.05$. Table 3.2 gives us values of e_α when we fix $p = 0.95$. Table 3.1 and Table 3.2 indicate that e_α is either equal to or very close to 1 in every case. Also note that $e_\alpha \le 1$ in every case. This implies that

Table 3.1 ($p = 0.05$, $n_1 = 25$, $n_2 = 30$, $a = 5$, $b = 10$, $c = 10$, $d = 5$)

	$e_{0.01}$	$e_{0.05}$	$e_{0.10}$	$e_{0.25}$	$e_{0.50}$
$\tau = 0.00$	1	1	1	1	1
$\tau = 0.05$	1	1	1	1	1
$\tau = 0.10$	1	1	1	1	1
$\tau = 0.15$	1	1	1	0.99950	0.99631
$\tau = 0.20$	1	1	0.99987	0.99518	0.99514
$\tau = 0.25$	1	0.99857	0.99558	0.99464	0.99659
$\tau = 0.30$	1	0.99513	0.99450	0.99592	0.99831
$\tau = 0.35$	0.99681	0.99457	0.99559	0.99793	0.99938
$\tau = 0.40$	0.99463	0.99586	0.99739	0.99916	0.99982
$\tau = 0.45$	0.99487	0.99764	0.99883	0.99974	0.99996
$\tau = 0.50$	0.99650	0.99898	0.99960	0.99994	0.99999
$\tau = 0.55$	0.99820	0.99967	0.99990	0.99999	0.99999
$\tau = 0.60$	0.99932	0.99992	0.99998	0.99999	1
$\tau = 0.65$	0.99982	0.99998	0.99999	1	1
$\tau = 0.70$	0.99997	0.99999	0.99999	1	1
$\tau = 0.75$	0.99999	1	1	1	1
$\tau = 0.80$	0.99999	1	1	1	1
$\tau = 0.85$	1	1	1	1	1
$\tau = 0.90$	1	1	1	1	1

Table 3.2 ($p = 0.95$, $n_1 = 25$, $n_2 = 30$, $a = 5$, $b = 10$, $c = 10$, $d = 5$)

	$e_{0.01}$	$e_{0.05}$	$e_{0.10}$	$e_{0.25}$	$e_{0.50}$
$\tau = 0.00$	1	1	1	1	1
$\tau = -0.05$	1	1	1	1	1
$\tau = -0.10$	1	1	1	1	1
$\tau = -0.15$	1	1	1	1	1
$\tau = -0.20$	1	1	1	1	1
$\tau = -0.25$	1	1	1	1	1
$\tau = -0.30$	1	1	1	1	1
$\tau = -0.35$	0.99999	0.99999	0.99999	0.99999	0.99999
$\tau = -0.40$	0.99999	0.99999	0.99999	0.99999	0.99999
$\tau = -0.45$	0.99999	0.99999	0.99999	0.99999	0.99999
$\tau = -0.50$	0.99999	0.99999	0.99999	0.99999	0.99999
$\tau = -0.55$	0.99998	0.99999	0.99999	0.99999	0.99999
$\tau = -0.60$	0.99998	0.99999	0.99999	0.99999	0.99999
$\tau = -0.65$	0.99998	0.99999	0.99999	0.99999	1
$\tau = -0.70$	0.99999	0.99999	0.99999	1	1
$\tau = -0.75$	0.99999	0.99999	0.99999	1	1
$\tau = -0.80$	0.99999	1	1	1	1
$\tau = -0.85$	1	1	1	1	1
$\tau = -0.90$	1	1	1	1	1

$R(\hat{p}^{(T)}_{1s\alpha}) \geq R(\hat{p}^{(T)}_{1N})$. The interpretation of these numerical results will be discussed in Section 4.7.

4.4 2 x s case.

As in Section 4.2, suppose we have two binomial populations, say π_1 and π_2. Also we have two kinds of individuals in each population, say α and β. Let $p_i = P(\alpha|\pi_i)$, $q_i = 1-p_i$, $0 < p_i < 1$, $i = 1,2$. Suppose we have the following past data available:

X_i α individuals out of total n_i observations from π_i, $i = 1,2$;

X_1 and X_2 are independent.

Suppose we have $s(s \geq 3)$ kinds of treatments, say $1,2,\ldots,s$.

Suppose Y_1 is a new individual from π_1. Y_1 is independent of X_i. We don't know if Y_1 is α or β. We want to make a treatment decision: apply j to Y_1, $j = 1,2,\ldots,s$. Suppose we have the following $2 \times s$ loss table.

	1	2		s-1	s
Y_1 being α	c_{11}	c_{12}		c_{1s-1}	c_{1s}
Y_1 being β	c_{21}	c_{22}		c_{2s-1}	c_{2s}

Table 4.1

Let

(4.1) Condition (i) $c_{1i_1} < c_{1i_2} < \ldots < c_{1i_{s-1}} < c_{1i_s}$

Condition (ii) $c_{2i_s} < c_{2i_{s-1}} < \ldots < c_{2i_2} < c_{2i_1}$,

where $(i_1,i_2,\ldots,i_s)$ is a permutation of $(1,2,\ldots,s)$.

Lemma 4.1. If the conditions (i) and (ii) do not hold, then the $2 \times s$ case will degenerate into the $2 \times \gamma$ case, $\gamma \leq s - 1$.

Proof: See Appendix C.

Because of Lemma 4.1, we can assume without loss of generality that

$$(4.2) \quad \begin{aligned} &(i) \quad c_{11} < c_{12} < \ldots < c_{1s-1} < c_{1s} \\ &(ii) \quad c_{2s} < c_{2s-1} < \ldots < c_{22} < c_{21} . \end{aligned}$$

It is easy to see that

(4.3) Expected loss when applying $j = p_1 c_{1j} + q_1 c_{2j}$.

We assume that each treatment is best for some range of p_1. Then we have the following result.

Lemma 4.2. Assume that for any j there is a p_1 (range of p_1) such that $p_1 c_{1j} + q_1 c_{2j} = \min_e (p_1 c_{1e} + q_1 c_{2e})$. Then

$$p_1 c_{1j} + q_1 c_{2j} \leq \min_e (p_1 c_{1e} + q_1 c_{2e})$$

if and only if

$$(4.4) \quad \begin{aligned} &(i) \quad \delta_j \leq p_1 \leq \delta_{j-1}, \quad \text{for} \quad j = 2,\ldots,s-1 , \\ &(ii) \quad p_1 \geq \delta_j, \quad \text{for} \quad j = 1 , \\ &(iii) \quad p_1 \leq \delta_{j-1}, \quad \text{for} \quad j = s , \end{aligned}$$

where

$$(4.5) \quad \delta_j = (c_{2j} - c_{2j+1})/(c_{1j+1} - c_{1j} + c_{2j} - c_{2j+1}).$$

Proof: See Appendix C.

It follows that $0 < \delta_{s-1} \leq \delta_{s-2} \leq \ldots \leq \delta_2 \leq \delta_1$. Without loss of generality we can assume that

$$(4.6) \qquad 0 < \delta_{s-1} < \delta_{s-2} < \ldots < \delta_2 < \delta_1 < 1 \ .$$

From (4.3), (4.4), and (4.6), we can define a treatment decision rule based on expected loss as follows:

$$(4.7) \quad \begin{cases} \text{Apply } j \text{ if } \delta_j < \tilde{p}_1 \leq \delta_{j-1} \text{ for } j = 2,\ldots,s-1 \\ \qquad\quad 1 \text{ if } \tilde{p}_1 > \delta_1 \\ \qquad\quad s \text{ if } \tilde{p}_1 \leq \delta_{s-1} \ , \end{cases}$$

where $\tilde{p}_1$ is an estimator of p_1 which is independent of Y_1.

Again, we define three treatment decision rules, namely the never-pool, the always-pool and the sometimes-pool as indicated in (2.9).

Lemma 4.3. The expected loss using $\tilde{p}_1$ is

$$(4.8) \qquad R(\tilde{p}_1) = c_{21} + (c_{11} - c_{21})p_1 - \sum_{j=1}^{s-1} d_j P(\tilde{p}_1 \leq \delta_j) \ ,$$

where

$$(4.9) \qquad d_j = (c_{2j} - c_{2j+1}) - (c_{1j+1} - c_{1j} + c_{2j} - c_{2j+1})p_1 \ .$$

Proof: See Appendix C.

Assume that n_1 and n_2 are sufficiently large such that approximately

$$(4.10) \qquad \hat{p}_{1N} \sim N(\mu_1, \sigma_1^2) \text{ and } \hat{p}_{1A} \sim N(\mu_2, \sigma_2^2) \ ,$$

where $\hat{p}_{1N}$ and $\hat{p}_{1A}$ are as indicated in (2.9) and (μ_i, σ_i^2) is as indicated in (2.12). We define approximate expected losses by

$$(4.11) \quad \begin{aligned} \tilde{R}(\hat{p}_{1N}) &= c_{21} + (c_{11} - c_{21})p_1 - \sum_{j=1}^{s-1} d_j \Phi[(\delta_j - \mu_1)/\sigma_1] \\ \tilde{R}(\hat{p}_{1A}) &= c_{21} + (c_{11} - c_{21})p_1 - \sum_{j=1}^{s-1} d_j \Phi[(\delta_j - \mu_2)/\sigma_2] \ . \end{aligned}$$

From (4.8) and (4.11), it follows that

(4.12) $\tilde{R}(\hat{p}_{1N}) \tilde{=} R(\hat{p}_{1N})$ and $\tilde{R}(\hat{p}_{1A}) \tilde{=} R(\hat{p}_{1A})$.

Theorem 4.1. A sufficient condition for $\tilde{R}(\hat{p}_{1A}) < \tilde{R}(\hat{p}_{1N})$ is that

(4.13) $(p_1,p_2) \in \cup_{k=1}^{s} A_k$,

where

$A_1 = \{0 < \mu_1 \leq \delta_{s-1} < 1$ and $\max_{I_1}[\delta_j(\sigma_2 - \sigma_1)] < \sigma_2\mu_1 - \sigma_1\mu_2\}$,

$A_k = \{0 < \delta_{s-k+1} < \mu_1 \leq \delta_{s-k} < 1$ and

$\max_{I_k}[\delta_j(\sigma_2-\sigma_1)] < \sigma_2\mu_1 - \sigma_1\mu_2 < \min_{J_k}[\delta_j(\sigma_2-\sigma_1)]\}$, k=2,...,s-1,

$A_s = \{\delta_1 < \mu_1 < 1$ and $\sigma_2\mu_1 - \sigma_1\mu_2 < \min_{I_1}[\delta_j(\sigma_2-\sigma_1)]\}$,

$I_1 = (j \mid j = 1,2,\ldots,s-1)$,

$I_k = (j \mid j = 1,2,\ldots,s-k)$,

$J_k = (j \mid j = s-k+1,\ldots,s-1)$.

Proof: From (4.11), it follows that

$$\tilde{R}(\hat{p}_{1A}) - \tilde{R}(\hat{p}_{1N}) = \sum_{j=1}^{s-1} d_j\{\Phi[(\delta_j-\mu_1)/\sigma_1] - \Phi[(\delta_j-\mu_2)/\sigma_2]\} .$$

Suppose $p_i \in A_k$. (i) $\delta_j(\sigma_2 - \sigma_1) < \sigma_2\mu_1 - \sigma_1\mu_2$, for all $j \in I_k$. It follows that $(\delta_j - \mu_1)/\sigma_1 < (\delta_j - \mu_2)/\sigma_2$, for all $j \in I_k$. Consequently, $\Phi[(\delta_j - \mu_1)/\sigma_1] < \Phi[(\delta_j - \mu_2)/\sigma_2]$, for all $j \in I_k$. (ii) $\sigma_2\mu_1 - \sigma_1\mu_2 < \delta_j(\sigma_2 - \sigma_1)$, for all $j \in J_k$. It follows that $\Phi[(\delta_j - \mu_1)/\sigma_1] > \Phi[(\delta_j - \mu_2)/\sigma_2]$ for all $j \in J_k$. (iii)

$\delta_{s-k+1} < \mu_1 \leq \delta_{s-k}$. It follows that $d_j < o$ for $j \in J_k$ and $d_j \geq 0$ for $j \in I_k$. The result follows immediately. Q.E.D.

Corollary 4.1. If $p_1 = p_2$, then $\tilde{R}(\hat{p}_{1A}) < \tilde{R}(\hat{p}_{1N})$.

Proof: It follows directly from Theorem 4.1.

Following the same ideas as we do in Sections 4.2 and 4.3, we can define a sometimes-pool treatment decision rule based on expected loss. Unfortunately Theorem 4.1 gives sufficient conditions only, not necessary and sufficient conditions. Since we are unable to find necessary and sufficient conditions, the following definition is based on the sufficient conditions.

Definition 4.1. We define a sometimes-pool treatment decision rule based on expected loss as follows:

$$(4.14) \qquad \begin{cases} \text{Apply } j \text{ if } \delta_j < \hat{p}_{1s} \leq \delta_{j-1} \text{ for } j = 2,\ldots,s-1 \\ \qquad\quad 1 \text{ if } \hat{p}_{1s} > \delta_1 \\ \qquad\quad s \text{ if } \hat{p}_{1s} \leq \delta_{s-1} , \end{cases}$$

where the pooling decision rule is as follows:

$$(4.15) \qquad \hat{p}_{1s} = \begin{cases} \hat{p}_{1A}, & \text{if } (\hat{p}_1, \hat{p}_2) \in \bigcup_{k=1}^{s} \hat{A}_k \\ \hat{p}_{1N}, & \text{otherwise} , \end{cases}$$

where

(4.16) $\hat{A}_k$ is the same as A_k in (4.13) except that we replace p_i by $\hat{p}_i$.

Lemma 4.4. The never-pool treatment decision rule and the always-pool treatment decision rule have different treatment decisions if

and only if

$$(4.17) \qquad (\hat{p}_1, \hat{p}_2) \in \hat{B}_1 \cup \hat{B}_2 ,$$

where

$$\hat{B}_1 = \{\hat{p}_{1N} \le \delta_j < \hat{p}_{1A}, \text{ for some } j = 1,2,\ldots,s-1\}$$

$$\hat{B}_2 = \{\hat{p}_{1A} \le \delta_j < \hat{p}_{1N}, \text{ for some } j = 1,2,\ldots,s-1\} .$$

Proof: It follows directly from (4.7) and (2.9).

Theorem 4.2. The never-pool treatment decision rule and the sometimes-pool treatment decision rule (defined in Definition 4.1) always make the same treatment decision.

Proof: Similar to the proof of Theorem 2.2.

Corollary 4.2. $R(\hat{p}_{1N}) = R(\hat{p}_{1s})$.

Proof: It follows directly from Theorem 4.2.

4.5. r x 2 case.

Suppose we have two r-variate ($r \ge 3$) multinomial populations, say π_1 and π_2. Also we have r kinds of individuals in each population, say α_j, $j = 1,2,\ldots,r$. Let

$$(5.1) \qquad \begin{aligned} &p_{ij} = P(\alpha_j|\pi_i), \quad i = 1,2, \quad j = 1,2,\ldots,r , \\ &0 < p_{ij} < 1, \quad \Sigma_j p_{ij} = 1, \quad i = 1,2. \end{aligned}$$

Suppose we have the following past data available:

$$(5.2) \qquad \begin{aligned} &X_{ij} \ \alpha_j \text{ individuals out of total } n_i \text{ observations from} \\ &\pi_i, \ i = 1,2; \\ &X_{1j} \text{ and } X_{2j}, \text{ are independent.} \end{aligned}$$

Suppose we have two kinds of treatment, say 1 and 2.

Suppose Y_1 is a new individual from π_1. Y_1 is independent of X_{ij}. We don't know if Y_1 is α_j or not. We want to make a treatment decision: apply 1 or 2 to Y_1. Suppose we have the following $r \times 2$ loss table.

	1	2
Y_1 being α_1	c_{11}	c_{12}
Y_1 being α_2	c_{21}	c_{22}
	⋮	⋮
Y_1 being α_{r-1}	c_{r-11}	c_{r-12}
Y_1 being α_r	c_{r1}	c_{r2}

Table 5.1

It follows that

$$(5.3)\quad \begin{cases} \text{Expected loss when applying } 1 = \sum_{j=1}^{r-1}(c_{j1} - c_{r1})p_{1j} + c_{r1} \\ \text{Expected loss when applying } 2 = \sum_{j=1}^{r-1}(c_{j2} - c_{r2})p_{1j} + c_{r2}. \end{cases}$$

Consequently,

$$(5.4)\quad \text{Expected loss when applying } 1 < \text{Expected loss when apply } 2 \Leftrightarrow \sum_{j=1}^{r-1} \lambda_j p_{1j} > \lambda ,$$

where

$$(5.5)\quad \begin{aligned} \lambda_j &= (c_{j2} - c_{r2}) - (c_{j1} - c_{r1}),\ j = 1,2,\ldots,r-1 , \\ \lambda &= c_{r1} - c_{r2} . \end{aligned}$$

Therefore, we can define a treatment decision rule based on expected loss as follows:

$$(5.6)\quad \begin{cases} \text{Apply } 1 \text{ if } \sum_{j=1}^{r-1} \lambda_j \tilde{p}_{1j} > \lambda \\ \text{Apply } 2 \text{ if } \sum_{j=1}^{r-1} \lambda_j \tilde{p}_{1j} \le \lambda , \end{cases}$$

where $\tilde{p}_{1j}$ is an estimator of p_{1j}. $\sum_{j=1}^{r} \tilde{p}_{1j} = 1$. $\tilde{p}_{1j}$ is independent of Y_1.

Let

$$(5.7)\quad \hat{p}_{ij} = X_{ij}/n_i,\ j = 1,2,\ldots,r,\quad i = 1,2.$$

We define three treatment decision rules as follows:

(i) Never-pool treatment decision rule: put $\tilde{p}_{1j} = \hat{p}_{1jN} = \hat{p}_{1j}$, for all j.

(ii) Always-pool treatment decision rule: put

$$(5.8)\quad \tilde{p}_{1j} = \hat{p}_{1jA} = \sum_{i=1}^{2} n_i \hat{p}_{ij}/(n_1 + n_2) \text{ for all } j.$$

(iii) Sometimes-pool treatment decision rule: put $\tilde{p}_{1j} = \hat{p}_{1js}$ for all j, where $\hat{p}_{1js}$ depends on the pooling decision rule which will be discussed later.

Lemma 5.1. The expected loss using $\tilde{p}_{1j}$ is

$$(5.9)\quad R(\tilde{p}_{1j}) = \sum_{j=1}^{r} c_{j1} p_{1j} + \Big(\sum_{j=1}^{r-1} \lambda_j p_{1j} - \lambda\Big) P\Big(\sum_{j=1}^{r-1} \lambda_j \tilde{p}_{1j} \le \lambda\Big).$$

Proof: See Appendix C.

Assume that n_1 and n_2 are sufficiently large such that approximately

$$(5.10)\quad \begin{aligned} \sum_{j=1}^{r-1} \lambda_j \hat{p}_{1jN} &\sim N(\mu_1, \sigma_1^2) \\ \sum_{j=1}^{r-1} \lambda_j \hat{p}_{1jA} &\sim N(\mu_2, \sigma_2^2) , \end{aligned}$$

where

$$\mu_1 = \sum_{j=1}^{r-1} \lambda_j p_{1j} ,$$

$$\sigma_1^2 = \sum_{j=1}^{r-1} \lambda_j^2 p_{1j}(1-p_{1j})/n_1 - 2\sum_{j<j'} \lambda_j \lambda_{j'} p_{1j} p_{1j'}/n_1 ,$$

$$\mu_2 = \sum_{j=1}^{r-1} \lambda_j \sum_{i=1}^{2} n_i p_{ij}/(n_1+n_2) , \tag{5.11}$$

$$\sigma_2^2 = \sum_{j=1}^{r-1} \lambda_j^2 \sum_{i=1}^{2} n_i p_{ij}(1-p_{ij})/(n_1+n_2)^2 - 2\sum_{j<j'} \lambda_j \lambda_{j'} \sum_{i=1}^{2} n_i p_{ij} p_{ij'}/(n_1+n_2)^2 .$$

We define approximate expected losses by

$$\begin{aligned} \tilde{R}(\hat{p}_{1jN}) &= \sum_{j=1}^{r} c_{j1} p_{1j} + (\sum_{j=1}^{r-1} \lambda_j p_{1j} - \lambda) \Phi[(\lambda-\mu_1)/\sigma_1] \\ \tilde{R}(\hat{p}_{1jA}) &= \sum_{j=1}^{r} c_{j1} p_{1j} + (\sum_{j=1}^{r-1} \lambda_j p_{1j} - \lambda) \Phi[(\lambda-\mu_2)/\sigma_2] . \end{aligned} \tag{5.12}$$

From (5.19) and (5.12), it follows that

$$\tilde{R}(\hat{p}_{1jN}) \cong R(\hat{p}_{1jN}) \quad \text{and} \quad \tilde{R}(\hat{p}_{1jA}) \cong R(\hat{p}_{1jA}) . \tag{5.13}$$

Theorem 5.1.

$$\tilde{R}(\hat{p}_{1jA}) < \tilde{R}(\hat{p}_{1jN}) \Leftrightarrow (p_{ij},\ i = 1,2,\ j = 1,2,\ldots,r-1) \in A_1 \cup A_2 , \tag{5.14}$$

where $A_1 = C \cap B$, $A_2 = \bar{C} \cap \bar{B}$, $C = (\mu_1 < \lambda)$, $\bar{C} = (\mu_1 > \lambda)$, $B = \{\sigma_2\mu_1 - \sigma_1\mu_2 > \lambda(\sigma_2 - \sigma_1)\}$ and $\bar{B} = \{\sigma_2\mu_1 - \sigma_1\mu_2 < \lambda(\sigma_2 - \sigma_1)\}$.

Proof: Similar to the proof of Theorem 2.1.

Corollary 5.1. If $p_{1j} = p_{2j}$, for all j, then $\tilde{R}(\hat{p}_{1jA}) < \tilde{R}(\hat{p}_{1jN})$.

Proof: It follows directly from Theorem 5.1.

Definition 5.1. We define a sometimes-pool treatment decision rule based on expected loss as follows:

$$(5.15)\quad \begin{cases} \text{Apply } 1 \text{ if } \sum_{j=1}^{r-1} \lambda_j \hat{p}_{1js} > \lambda \\ \text{Apply } 2 \text{ if } \sum_{j=1}^{r-1} \lambda_j \hat{p}_{1js} \le \lambda , \end{cases}$$

where the pooling decision rule is as follows:

$$(5.16)\quad \hat{p}_{1js} = \begin{cases} \hat{p}_{1jA} & \text{if } \hat{p}_{ij} \in \hat{A}_1 \cup \hat{A}_2 \text{ for } i = 1,2 \\ \hat{p}_{1jN} & \text{otherwise} \end{cases}$$

for all j, where

(5.17) $\hat{A}_i$ is the same as A_i in (5.14) except that we replace p_{ij} by $\hat{p}_{ij}$.

Lemma 5.2. The never-pool treatment decision rule and the always-pool treatment decision rule have different treatment decisions if and only if

$$(5.18)\quad (\hat{p}_{ij},\ i = 1,2,\ j = 1,2,\ldots,r-1) \in \hat{B}_1 \cup \hat{B}_2 ,$$

where

$$\hat{B}_1 = \{\sum_{j=1}^{r-1} \lambda_j \hat{p}_{1jN} \le \lambda < \sum_{j=1}^{r-1} \lambda_j \hat{p}_{1jA}\}$$

$$\hat{B}_2 = \{\sum_{j=1}^{r-1} \lambda_j \hat{p}_{1jA} \le \lambda < \sum_{j=1}^{r-1} \lambda_j \hat{p}_{1jN}\} .$$

Proof: It follows directly from (5.6) and (5.8).

Theorem 5.2. The never-pool treatment decision rule and the sometimes-pool treatment decision rule (defined in Definition 5.1) always make the same treatment decision.

Proof: Similar to the proof of Theorem 2.2.

Corollary 5.2. $R(\hat{p}_{1jN}) = R(\hat{p}_{1js})$.

Proof: It follows directly from Theorem 5.2.

4.6 Prediction decisions in the two normal populations case.

Suppose we have two normal populations, say $N_1(\mu_1,\sigma_1^2)$ and $N_2(\mu_2,\sigma_2^2)$. Suppose we have the following past data available: n_i observations $(Y_{ij}, j = 1,2,\ldots,n_i)$ from $N_i(\mu_i,\sigma_i^2)$, $i = 1,2$. Y_{ij}'s are independent. Let

$$(6.1)\quad \hat{\mu}_i = \bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i, \quad \hat{\sigma}_i^2 = \sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2/(n_i-1), \quad i = 1,2.$$

Suppose we have a new observation , say Y_1, from $N_1(\mu_1,\sigma_1^2)$. Y_1 is independent of Y_{ij}. We want to predict if $Y_1 \le c$ or $Y_1 > c$ (where c is a real number).

We can define a prediction decision rule as follows:

$$(6.2)\quad \begin{aligned} &\text{(i)} \quad \text{Predict that } Y_1 \le c \text{ if } \tilde{\mu}_1 \le c \\ &\text{(ii)} \quad \text{Predict that } Y_1 > c \text{ if } \tilde{\mu}_1 > c \ , \end{aligned}$$

where $\tilde{\mu}_1$ is an estimator of μ_1. $\tilde{\mu}_1$ is independent of Y_1. Then, we are subject to two kinds of error:

(i) Predict that $Y_1 \le c$ when in fact $Y_1 > c$

(ii) Predict that $Y_1 > c$ when in fact $Y_1 \le c$.

Hence we have the following loss table:

	$Y_1 \le c$	$Y_1 > c$
$\tilde{\mu}_1 \le c$	0	$a > 0$
$\tilde{\mu}_1 > c$	$b > 0$	0

Table 6.1

We define three prediction decision rules as follows:

(6.3)

(i) Never-pool prediction decision rule: put $\tilde{\mu}_1 = \hat{\mu}_{1N} = \hat{\mu}_1$.

(ii) Always-pool prediction decision rule: put $\tilde{\mu}_1 = \hat{\mu}_{1A} = \sum_{i=1}^{2} n_i \hat{\mu}_i / (n_1 + n_2)$.

(iii) Sometimes-pool prediction decision rule: put $\tilde{\mu}_1 = \hat{\mu}_{1s}$, where $\hat{\mu}_{1s}$ depends on the pooling decision rule.

Lemma 6.1. The expected loss using $\tilde{\mu}_1$ is

$$(6.4) \quad R(\tilde{\mu}_1) = bP(Y_1 \leq c) + [a-(a+b)P(Y_1 \leq c)]P(\tilde{\mu}_1 \leq c) .$$

Proof: See Appendix C.

Let

$$(6.5) \quad \mu_1^* = \mu_1, \ \mu_2^* = \sum_{i=1}^{2} n_i \mu_i / (n_1 + n_2), \ \sigma_1^{*2} = \sigma_1^2 / n_1 \text{ and } \sigma_2^{*2} = \sum_{i=1}^{2} n_i \sigma_i^2 / (n_1 + n_2)^2 .$$

Theorem 6.1.

$$(6.6) \quad R(\hat{\mu}_{1A}) < R(\hat{\mu}_{1N}) \Leftrightarrow (\mu_i, \sigma_i^2 \ i = 1,2) \in A_1 \cup A_2 ,$$

where $A_1 = C \cap B$, $A_2 = \bar{C} \cap \bar{B}$, $C = \{(c-\mu_1)/\sigma_1 > \Phi^{-1}[a/(a+b)]\}$, $\bar{C} = \{(c-\mu_1)/\sigma_1 < \Phi^{-1}[a/(a+b)]\}$, $B = \{\sigma_2^* \mu_1^* - \sigma_1^* \mu_2^* > c(\sigma_2^* - \sigma_1^*)\}$ and $\bar{B} = \{\sigma_2^* \mu_1^* - \sigma_1^* \mu_2^* < c(\sigma_2^* - \sigma_1^*)\}$.

Proof: Similar to the proof of Theorem 2.1.

Corollary 6.1. If $a = b$, $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$, then $R(\hat{\mu}_{1A}) < R(\hat{\mu}_{1N})$.

Proof: Note that $\Phi^{-1}[a/(a+b)] = 0$ when $a = b$. The result follows immediately from Theorem 6.1. Q.E.D.

As we do in previous sections, we can define a sometimes-pool prediction decision rule based on expected loss.

Definition 6.1. We define a sometimes-pool prediction decision rule based on expected loss as follows:

$$(6.7)\quad \begin{cases} \text{Predict that } Y_1 \le c \text{ if } \hat{\mu}_{1s} \le c \\ \text{Predict that } Y_1 > c, \text{ if } \hat{\mu}_{1s} > c \ , \end{cases}$$

where the pooling decision rule is as follows:

$$(6.8)\quad \hat{\mu}_{1s} = \begin{cases} \hat{\mu}_{1A}, & \text{if } (\hat{\mu}_i, \hat{\sigma}_i^2, \ i = 1,2) \in \hat{A}_1 \cup \hat{A}_2 \\ \hat{\mu}_{1N}, & \text{otherwise} \ , \end{cases}$$

where

(6.9) $\hat{A}_i$ is the same as A_i in (6.6) except that we replace (μ_i, σ_i^2) by $(\hat{\mu}_i, \hat{\sigma}_i^2)$.

Lemma 6.2. The never-pool prediction decision rule and the always-pool prediction decision rule have different prediction decisions if and only if

$$(6.10)\quad (\hat{\mu}_i, \hat{\sigma}_i^2, \ i = 2) \in \hat{B}_1 \cup \hat{B}_2 \ ,$$

where

$$\hat{B}_1 = (\hat{\mu}_{1N} \le c < \hat{\mu}_{1A}) \quad \text{and} \quad \hat{B}_2 = (\hat{\mu}_{1A} \le c < \hat{\mu}_{1N}) \ .$$

Proof: It follows directly from (6.2) and (6.3).

Theorem 6.2. Suppose $a = b$. Then the never-pool prediction decision rule and the sometimes-pool prediction decision rule (defined in Definition 6.1) always makes the same prediction decision.

Proof: Note that $\xi^{-1}(a/(a+b))= 0$ when $a = b$. The rest of the arguments are similar to the proof of Theorem 2.2. Q.E.D.

Corollary 6.2. Suppose $a = b$. Then $R(\hat{\mu}_{1N}) = R(\hat{\mu}_{1s})$.

Proof: It follows directly from Theorem 6.2.

4.7. Discussion

We have discussed the problem of pooling data from a decision theoretic point of view. We find that the result is a bit surprising. We define the never-pool treatment (prediction) decision rule (when using the never-pooled estimator) and the always-pool treatment (prediction) decision rule (when using the always-pooled estimator). Then we find the set of parameters for which the always-pool treatment (prediction) decision rule is better than the never-pool. We call this set of parameters the "pooling region". Since parameters are unknown, we replace parameters by their respective estimators. Consequently, we have a pooling decision rule based on expected loss as follows: poool data if estimators fall in the "pooling region"; don't pool, otherwise. Then we have a sometimes-pool treatment (prediction) decision rule based on this pooling decsision rule. This is justified intuitively.

However, the sometimes-pool treatment (prediction) decision rule based on this pooling decision rule always makes the same treatment (prediction) decision as the never-pool does. Hence their expected losses are equal. The reason for this phenomenon is that the additional data do not change our treatment (prediction) decision when we pool the data.

Another intuitively appealing pooling decision rule is the one based on a preliminary test. In Section 4.3 we have studied the sometimes-pool treatment decision rule based on this rule. The numerical study in Section 4.3 indicates that this sometimes-pool treatment decision rule is worse for most parameter values than the never-pool, as far as expected loss is concerned. Consequently, this implies that the pooling decision rule based on a preliminary test is usually worse than the pooling decision rule based on expected loss.

Therefore, it is better to use the pooling decision rule based on expected loss rather than based on a preliminary test. But the pooling decision rule based on expected loss does not affect our treatment decisions. Hence pooling of data is irrelevant in this sense.

CHAPTER 5

A DECISION THEORETIC APPROACH TO THE PROBLEM OF POOLING DATA (THREE POPULATIONS CASE)

5.1. Introduction.

In Chapter 4, we have shown that the sometimes-pool treatment decision rule based on expected loss and the never-pool treatment decision rule always make the same treatment decision in the two populations case. In this chapter, we generalize this result in the three populations case. The discussions are mainly analogous to the two populations case. Sections 5.2 to 5.6 correspond to Sections 4.2 to 4.6.

Since most arguments are very similar, we give detailed discussions only in Section 5.2. For other sections, we simply describe the way we get results without going into details.

5.2. 2 X 2 case.

Let π_1, π_2, π_3 denote three binomial populations and let $\alpha, \beta, p_i, n_i, X_i, Y_1, a, b, c, d$ have the same meaning as in Section 4.2.

Let $\hat{p}_i = X_i/n_i$, $i = 1,2,3$. We define five treatment decision rules as follows:

(i) Never-pool treatment decision rule: put $\tilde{p}_1 = \hat{p}_{1N} = \hat{p}_1$. (2.1)

(ii) Always-pool treatment decision rule (I): put

$$\tilde{p}_1 = \hat{p}_{1A}^{(1)} = (n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1 + n_2).$$

(iii) Always-pool treatment decision rule (II): put

$$\tilde{p}_1 = \hat{p}_{1A}^{(2)} = (n_1\hat{p}_1 + n_3\hat{p}_3)/(n_1 + n_3).$$

(iv) Always-pool treatment decision rule (III): put

$$\tilde{p}_1 = \hat{p}_{1A}^{(3)} = \Sigma n_i \hat{p}_i / (\Sigma n_i).$$

(v) Sometimes-pool treatment decision rule: put $\tilde{p}_1 = \hat{p}_{1S}$, where $\hat{p}_{1s}$ depends on the pooling decision rule which will be discussed later.

Lemma 2.1, Section 4.2, continues to hold.

Assume that n_i, $i = 1,2,3$, are sufficiently large such that approximately

$$(2.2) \qquad \hat{p}_{1N} \sim N(\mu_0,\sigma_0^2) \quad \text{and} \quad \hat{p}_{1A}^{(j)} \sim N(\mu_j,\sigma_j^2), \quad j = 1,2,3,$$

where

$$\mu_0 = p_1,\ \mu_1 = (n_1p_1+n_2p_2)/(n_1+n_2),\ \mu_2 = (n_1p_1+n_3p_3)/(n_1+n_3),$$

$$\mu_3 = (\Sigma n_ip_i)/(\Sigma n_i),\ \sigma_0^2 = [p_1(1-p_1)]/n_1,\ \sigma_1^2 = [n_1p_1(1-p_1)$$

$$(2.3) \qquad +n_2p_2(1-p_2)]/(n_1+n_2)^2,$$

$$\sigma_2^2 = [n_1p_1(1-p_1)+n_3p_3(1-p_3)]/(n_1+n_3)^2,\ \sigma_3^2 = [\Sigma n_ip_i(1-p_i)]/(\Sigma n_i)^2$$

We define approximate expected losses by

$$(2.4) \qquad \begin{aligned} &(i)\ \tilde{R}(\hat{p}_{1N}) = c + (a-c)p_1 + \{[(b-a)+(c-d)]p_1-(c-d)\}\Phi[(\delta-\mu_0)/\sigma_0], \\ &(ii)\ \tilde{R}(\hat{p}_{1A}^{(j)}) = c + (a-c)p_1+\{[(b-a)+(c-d)]p_1-(c-d)\}\Phi[(\delta-\mu_j)/\sigma_j] \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j = 1,2,3, \end{aligned}$$

where Φ stands for the standard normal c.d.f.. From Lemma 2.1 (Section 4.2) and (2.4), it follows that

$$(2.5) \qquad \tilde{R}(\hat{p}_{1N}) \tilde{=} R(\hat{p}_{1N}) \quad \text{and} \quad \tilde{R}(\hat{p}_{1A}^{(j)}) \tilde{=} R(\hat{p}_{1A}^{(j)}), \quad j = 1,2,3.$$

Theorem 2.1.

$$(i)\ \tilde{R}(\hat{p}_{1A}^{(1)}) < \min[\tilde{R}(\hat{p}_{1N}),\tilde{R}(\hat{p}_{1A}^{(2)}),\tilde{R}(\hat{p}_{1A}^{(3)})] \Leftrightarrow (p_i,\ i = 1,2,3) \in A_{11} \cap A_{12}$$

(2.6) (ii) $\tilde{R}(\hat{p}_{1A}^{(2)}) < \min[\tilde{R}(\hat{p}_{1N}), \tilde{R}(\hat{p}_{1A}^{(1)}), \tilde{R}(\hat{p}_{1A}^{(3)})] \Leftrightarrow (p_i,\ i=1,2,3) \in A_{21} \cup A_{22}$

(iii) $\tilde{R}(\hat{p}_{1A}^{(3)}) < \min[\tilde{R}(\hat{p}_{1N}), \tilde{R}(\hat{p}_{1A}^{(1)}), \tilde{R}(\hat{p}_{1A}^{(2)})] \Leftrightarrow (p_i, i=1,2,3) \in A_{31} \cup A_{32}$,

where $A_{11} = C \cap B_{10} \cap B_{12} \cap B_{13}$, $A_{12} = \bar{C} \cap \bar{B}_{10} \cap \bar{B}_{12} \cap \bar{B}_{13}$, $A_{21} = C \cap B_{20} \cap B_{21} \cap B_{23}$, $A_{22} = \bar{C} \cap \bar{B}_{20} \cap \bar{B}_{21} \cap \bar{B}_{23}$, $A_{31} = C \cap B_{30} \cap B_{31} \cap B_{32}$, $A_{32} = \bar{C} \cap \bar{B}_{30} \cap \bar{B}_{31} \cap \bar{B}_{32}$, $C = (\mu_0 < \delta)$, $\bar{C} = (\mu_0 > \delta)$, $B_{ij} = \{\sigma_i\mu_j - \sigma_j\mu_i > \delta(\sigma_i - \sigma_j)\}$ and $\bar{B}_{ij} = \{\sigma_i\mu_j - \sigma_j\mu_i < \delta(\sigma_i - \sigma_j)\}$.

Proof: To prove (i), by (2.4),

$$\tilde{R}(\hat{p}_{1A}^{(1)}) - \tilde{R}(\hat{p}_{1N}) = \{[(b-a)+(c-d)]p_1 - (c-d)\}\{\Phi[(\delta-\mu_1)/\sigma_1] - \Phi[(\delta-\mu_0)/\sigma_0]\} .$$

It follows that $\tilde{R}(\hat{p}_{1A}^{(1)}) < \tilde{R}(\hat{p}_{1N}) \Leftrightarrow$

either (i) $\mu_0 < \delta$ and $\Phi[(\delta-\mu_1)/\sigma_1] > \Phi[(\delta-\mu_0)/\sigma_0]$,

or (ii) $\mu_0 > \delta$ and $\Phi[(\delta-\mu_1)/\sigma_1] < \Phi[(\delta-\mu_0)/\sigma_0]$.

But $\Phi[(\delta-\mu_1)/\sigma_1] > \Phi[(\delta-\mu_0)/\sigma_0] \Leftrightarrow \sigma_1\mu_0 - \sigma_0\mu_1 > \delta(\sigma_1-\sigma_0)$. Hence $\tilde{R}(\hat{p}_{1A}^{(1)}) < \tilde{R}(\hat{p}_{1N}) \Leftrightarrow$

(2.7) either (i) $\mu_0 < \delta$ and $\sigma_1\mu_0 - \sigma_0\mu_1 > \delta(\sigma_1 - \sigma_0)$,

or (ii) $\mu_0 > \delta$ and $\sigma_1\mu_0 - \sigma_0\mu_1 < \delta(\sigma_1 - \sigma_0)$.

Similarly, $\tilde{R}(\hat{p}_{1A}^{(1)}) < \tilde{R}(\hat{p}_{1A}^{(2)}) \Leftrightarrow$

(2.8) either (i) $\mu_0 < \delta$ and $\sigma_1\mu_2 - \sigma_2\mu_1 > \delta(\sigma_1 - \sigma_2)$,

or (ii) $\mu_0 > \delta$ and $\sigma_1\mu_2 - \sigma_2\mu_1 < \delta(\sigma_1 - \sigma_2)$.

Also, $\tilde{R}(\hat{p}_{1A}^{(1)}) < \tilde{R}(\hat{p}_{1A}^{(3)}) \Leftrightarrow$

(2.9) either (i) $\mu_0 < \delta$ and $\sigma_1\mu_3 - \sigma_3\mu_1 > \delta(\sigma_1 - \sigma_3)$,

or (ii) $\mu_0 > \delta$ and $\sigma_1\mu_3 \quad \sigma_3\mu_1 < \delta(\sigma_1 - \sigma_3)$.

(i) in (2.6) follows from (2.7), (2.8) and (2.9). Similar arguments give (ii) and (iii) in (2.6). Q.E.D.

Corollary 2.1. If $p_1 = p_2 = p_3$, then
$\tilde{R}(\hat{p}_{1A}^{(3)}) < \min[\tilde{R}(\hat{p}_{1N}), \tilde{R}(\hat{p}_{1A}^{(1)}), \tilde{R}(\hat{p}_{1A}^{(2)})]$.

Proof: It follows directly from Theorem 2.1.

Definition 2.1. We define a sometimes-pool treatment decision rule based on expected loss as follows:

$$(2.10) \quad \begin{cases} \text{Treat if} & \hat{p}_{1s} > \delta \\ \text{Don't treat is} & \hat{p}_{1s} \leq \delta \ , \end{cases}$$

where the pooling decision rule is as follows:

$$(2.11) \quad \hat{p}_{1s} = \begin{cases} \hat{p}_{1A}^{(1)}, & \text{if } (\hat{p}_i,\ i = 1,2,3) \in \hat{A}_{11} \cup \hat{A}_{12} \\ \hat{p}_{1A}^{(2)}, & \text{if } (\hat{p}_i,\ i = 1,2,3) \in \hat{A}_{21} \cup \hat{A}_{22} \\ \hat{p}_{1A}^{(3)}, & \text{if } (\hat{p}_i,\ i = 1,2,3) \in \hat{A}_{31} \cup \hat{A}_{32} \\ \hat{p}_{1N}\ , & \text{otherwise} \ , \end{cases}$$

where

(2.12) $\hat{A}_{ij}$ is the same as A_{ij} in (2.6) except that we replace (μ_i,σ_i) by $(\hat{\mu}_i,\hat{\sigma}_i)$; while $\hat{\mu}_i$ and $\hat{\sigma}_i$ are obtained from μ_i and σ_i in (2.3) by replacing p_i by $\hat{p}_i$.

Lemma 2.1. The never-pool and the always-pool treatment decision rules I [II, III] have different treatment decisions if and only if

$$(2.13) \quad \hat{p}_i \in \hat{B}_1[\hat{B}_2, \hat{B}_3] \quad \text{for} \quad i = 1,2,3 \ ,$$

where $\hat{B}_j = \{\hat{p}_{1N} \leq \delta < \hat{p}_{1A}^{(j)}\} \cup \{\hat{p}_{1A}^{(j)} \leq \delta < \hat{p}_{1N}\}$.

Proof: It follows directly from (2.6) (Section 4.2) and (2.1).

Remark: In the definition of $\hat{B}_j$ we could also write $\hat{\mu}_0$ and $\hat{\mu}_j$ in place of $\hat{p}_{1N}$ and $\hat{p}_{1A}^{(j)}$.

Theorem 2.2. The never-pool treatment decision rule and the sometimes-pool treatment decision rule defined above always make the same treatment decision.

Proof: Let $\hat{A}_j = \hat{A}_{j1} \cup \hat{A}_{j2}$, $j = 1,2,3$. By (2.11) and (2.13), it is sufficient to show that $\hat{A}_j \subset \hat{B}_j^c$ for $j = 1,2,3$. This is equivalent to showing that $\hat{B}_j \subset \hat{A}_j^c$ for $j = 1,2,3$. First to show that $\hat{B}_1 \subset \hat{A}_1^c$. Suppose $\hat{\mu}_0 \leq \delta < \hat{\mu}_1$. Case (i) $\hat{\mu}_0 = \delta$. It is clear that $(\hat{p}_i,\ i = 1,2,3) \in \hat{A}_1^c$. Case (ii) $\hat{\mu}_0 < \delta < \hat{\mu}_1$. Then $\hat{\sigma}_1(\hat{\mu}_0 - \delta) < \hat{\sigma}_0(\hat{\mu}_1 - \delta)$. It follows that $\hat{\sigma}_1\hat{\mu}_0 - \hat{\sigma}_0\hat{\mu}_1 < \delta(\hat{\sigma}_1 - \hat{\sigma}_0)$. Therefore $(\hat{p}_i,\ i=1,2,3) \in \hat{A}_1^c$. Similar arguments for $\hat{\mu}_1 \leq \delta < \hat{\mu}_0$. We have shown that $\hat{B}_1 \subset \hat{A}_1^c$. Similarly we can show that $\hat{B}_2 \subset \hat{A}_2^c$ and $\hat{B}_3 \subset \hat{A}_3^c$. Q.E.D.

Corollary 2.2. $R(\hat{p}_{1N}) = R(\hat{p}_{1s})$.

Proof: It follows directly from Theorem 2.2.

5.3. Arcsine square root transformation in the 2 X 2 case.

Assume that the formulation of problem and assumptions are the same as in Section 4.3 except that we have three populations now instead of two populations.

In a similar fashion as in Section 5.2, we define five treatment decision rules as follows:

(i) Never-pool treatment decision rule: put $\tilde{p}_1^{(T)} = \hat{p}_{1N}^{(T)}$,

(ii) Always-pool treatment decision rule (I): put $\tilde{p}_1^{(T)} = \hat{p}_{1A}^{(T)}(1)$,

(iii) Always-pool treatment decision rule (II): put $\tilde{p}_1^{(T)} = \hat{p}_{1A}^{(T)}(2)$,

(iv) Always-pool treatment decision rule (III): put $\tilde{p}_1^{(T)} = \hat{p}_{1A}^{(T)}(3)$,

(v) Sometimes-pool treatment decision rule: put $\tilde{p}_1^{(T)} = \hat{p}_{1s}^{(T)}$, where $\hat{p}_{1s}^{(T)}$ depends on a pooling decision rule,

where $\hat{p}_{1N}^{(T)}$ and $\hat{p}_{1A}^{(T)}(j)$ are respectively the same as $\hat{p}_{1N}$ and $\hat{p}_{1A}^{(j)}$ in (2.1) except that we replace $\hat{p}_i$ by $\sin^{-1}\sqrt{\hat{p}_i}$. We note that the always-pool treatment decision rule (I) is the same as the always-pool treatment decision rule discussed in Section 4.3.

Again, we assume that n_i, $i = 1,2,3$, are sufficiently large such that approximately $\hat{p}_{1N}^{(T)}$ and $\hat{p}_{1A}^{(T)}(j)$ are distributed as normal distributions. Then based on Lemma 3.1 (Section 4.3), we can define approximate expected losses $\tilde{R}(\hat{p}_{1N}^{(T)})$ and $\tilde{R}(\hat{p}_{1A}^{(T)}(j))$ in a similar fashion as before.

We can find a necessary and sufficient condition such that $\tilde{R}(\hat{p}_{1A(i)}^{(T)}) < \min[\tilde{R}(\hat{p}_{1N}^{(T)}), \tilde{R}(\hat{p}_{1A(j)}^{(T)}), \tilde{R}(\hat{p}_{1A(k)}^{(T)})]$, $i = 1,2,3$. Based on this necessary and sufficient condition, we can define a sometimes-pool treatment decision rule based on expected loss in the same way as we do in Definition 2.1. Following the analogous arguments as in Section 5.2, we can have the following result: the never-pool treatment decision rule and the sometimes-pool treatment decision rule based on expected loss always make the same treatment decision.

5.4. 2 x s case.

Again, we assume that the formulation of problem and assumptions are the same as in Section 4.4. except that we have three populations now.

In this case, we can also define five treatment decision rules, namely, the never-pool, the always-pool I [II, III] and the sometimes-pool as indicated in (2.1).

Assume that n_i, $i = 1,2,3$, are sufficiently large such that (2.2) holds.

Based on Lemma 4.3 (Section 4.4), we define approximate expected losses by

$$(4.1)\quad \begin{aligned} &(i)\ \tilde{R}(\hat{p}_{1N}) = c_{21} + (c_{11} - c_{21})p_1 - \sum_{\ell=1}^{s-1} d_\ell \Phi[(\delta_\ell - \mu_0)/\sigma_0] \\ &(ii)\ \tilde{R}(\hat{p}_{1A}^{(j)}) = c_{21} + (c_{11} - c_{21})p_1 - \sum_{\ell=1}^{s-1} d_\ell \Phi[(\delta_\ell - \mu_j)/\sigma_j] , \\ &\qquad\qquad j = 1,2,3 . \end{aligned}$$

The following notations will be used in the statement of the following theorem:

$$C_1 = \{0 < \mu_0 \le \delta_{s-1} < 1\} ,$$

$$C_k = \{0 < \delta_{s-k+1} < \mu_0 \le \delta_{s-k} < 1\} ,$$

$$C_s = \{\delta_1 < \mu_0 < 1\} ,$$

$$B_{ij}^{(1)} = \{\max_{I_1}[\delta_e(\sigma_i - \sigma_j)] < \sigma_i\mu_j - \sigma_j\mu_i] ,$$

$$B_{ij}^{(k)} = \{\max_{I_k}[\delta_e(\sigma_i - \sigma_j)] < \sigma_i\mu_j - \sigma_j\mu_i < \min_{J_k}[\delta_e(\sigma_i - \sigma_j)]\} , \quad k = 2,3,\ldots,s-1 ,$$

$$B_{ij}^{(s)} = \{\sigma_i\mu_j - \sigma_j\mu_i < \min_{I_1}[\delta_e(\sigma_i - \sigma_j)]\} ,$$

where I_1, I_k and J_k are indicated in Theorem 4.1 (Section 4.4) except that we replace j by e.

Theorem 4.1.

$$(4.2)\quad \begin{aligned} &(i)\ \text{A sufficient condition for } \tilde{R}(\hat{p}_{1A}^{(1)}) < \min[\tilde{R}(\hat{p}_{1N}), \\ &\qquad \tilde{R}(\hat{p}_{1A}^{(2)}), \tilde{R}(\hat{p}_{1A}^{(3)})] \text{ is that } (p_i,\ i = 1,2,3) \in U_{k=1}^{s} A_{1k}, \\ &(ii)\ \text{A sufficient condition for } \tilde{R}(\hat{p}_{1A}^{(2)}) < \min[\tilde{R}(\hat{p}_{1N}), \\ &\qquad \tilde{R}(\hat{p}_{1A}^{(1)}), \tilde{R}(\hat{p}_{1A}^{(3)})] \text{ is that } (p_i,\ i = 1,2,3) \in U_{k=1}^{s} A_{2k} , \end{aligned}$$

(iii) A sufficient condition for $\tilde{R}(\hat{p}_{1A}^{(3)}) < \min[\tilde{R}(\hat{p}_{1N}), \tilde{R}(\hat{p}_{1A}^{(1)}), \tilde{R}(\hat{p}_{1A}^{(2)})]$ is that $(p_i, i = 1,2,3) \in U_{k=1}^{s} A_{3k}$,

where

$$A_{1k} = C_k \cap B_{10}^{(k)} \cap B_{12}^{(k)} \cap B_{13}^{(k)}$$

$$A_{2k} = C_k \cap B_{20}^{(k)} \cap B_{21}^{(k)} \cap B_{23}^{(k)}$$

$$A_{3k} = C_k \cap B_{30}^{(k)} \cap B_{31}^{(k)} \cap B_{32}^{(k)}.$$

Proof: Similar to the proof of Theorem 4.1 (Section 4.4).

As unfortunately as Theorem 4.1 in Section 4.4 Theorem 4.1 gives us only a sufficient condition, not a necessary and sufficient condition. Based on this sufficient condition, we can define a pooling decision rule as follows:

$$(4.3)\quad \hat{p}_{1s} = \begin{cases} \hat{p}_{1A}^{(1)}, & \text{if } (\hat{p}_i, i = 1,2,3) \in U_{k=1}^{s} \hat{A}_{1k} \\ \hat{p}_{1A}^{(2)}, & \text{if } (\hat{p}_i, i = 1,2,3) \in U_{k=1}^{s} \hat{A}_{2k} \\ \hat{p}_{1A}^{(3)}, & \text{if } (\hat{p}_i, i = 1,2,3) \in U_{k=1}^{s} \hat{A}_{3k} \\ \hat{p}_{1N}, & \text{otherwise}, \end{cases}$$

where

(4.4) $\hat{A}_{ik}$ is the same as A_{ik} in (4.2) except that we replace p_i by $\hat{p}_i$.

Based on this pooling decision rule, we define a sometimes-pool treatment decision rule based on expected loss as indicated in (4.14) (Section 4.4) except that $\hat{p}_{1s}$ is indicated in (4.3).

Then, we have the following result.

Theorem 4.2. The never-pool treatment decision rule and the sometimes-pool treatment decision rule based on expected loss always make the same treatment decision.

Proof: Similar to the proof of Theorem 2.2.

5.5 r x 2 case.

Assume that the formulation of problem and assumptions are the same as in Section 4.5 except that we have three populations now.

Let $\hat{p}_{ij} = X_{ij}|n_i$, $i = 1,2,3$, $j = 1,2,\ldots,r$. We define five treatment decision rules as follows:

(i) Never-pool treatment decision rule: put $\tilde{p}_{1j} = \hat{p}_{1jN}$, for all j,

(ii) Always-pool treatment decision rule (I): put $\tilde{p}_{1j} = \hat{p}^{(1)}_{1jA}$, for all j,

(iii) Always-pool treatment decision rule (II): put $\tilde{p}_{1j} = \hat{p}^{(2)}_{1jA}$, for all j,

(iv) Always-pool treatment decision rule (III): put $\tilde{p}_{1j} = \hat{p}^{(3)}_{1jA}$, for all j,

(v) Sometimes-pool treatment decision rule: put $\tilde{p}_{1j} = \hat{p}_{1js}$, for all j, where $\hat{p}_{1js}$ depends on a pooling decision rule,

where $\hat{p}_{1jN}$ and $\hat{p}^{(k)}_{1jA}$ are the same as $\hat{p}_{1N}$ and $\hat{p}^{(k)}_{1A}$ in (2.1) except that we replace $\hat{p}_i$ by $\hat{p}_{ij}$.

Assume that n_i, $i = 1,2,3$, are sufficiently large such that approximately $\sum_{j=1}^{r-1} \lambda_j \hat{p}_{1jN}$ and $\sum_{j=1}^{r-1} \lambda_j \hat{p}^{(k)}_{1jA}$ are distributed as normal distributions. By Lemma 5.1 (Section 4.5) and normal approximations, we can define approximate expected losses $\tilde{R}(\hat{p}_{1jN})$ and $\tilde{R}(\hat{p}^{(k)}_{1jA})$. Then we can find a necessary and sufficient condition for which $\tilde{R}(\hat{p}^{(k)}_{1jA}) < \min[\tilde{R}(\hat{p}_{1jN}), \tilde{R}(\hat{p}^{(m)}_{1jA}), \tilde{R}(\hat{p}^{(n)}_{1jA})]$, $k = 1,2,3$. Based on this necessary and sufficient condition, we dan define a sometimes-pool treatment decision rule based on expected loss in a

similar fashion as in Definition 2.1. Then we can have the following result: the never-pool treatment decision rule and the sometimes-pool treatment decision rule based on expected loss always make the same treatment decision.

5.6. Prediction decisions in the three normal populations case.

Assume that the formulation of problem and assumptions are the same as in Section 4.6 except that we have three populations now. Let $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ have the same meaning as in Section 4.6. We define five prediction decision rules as follows:

(i) Never-pool prediction decision rule: put $\tilde{\mu}_1 = \hat{\mu}_{1N}$,

(ii) Always-pool prediction decision rule (I): put $\tilde{\mu}_1 = \hat{\mu}_{1A}^{(1)}$,

(iii) Always-pool prediction decision rule (II): put $\tilde{\mu}_1 = \hat{\mu}_{1A}^{(2)}$,

(iv) Always-pool prediction decision rule (III): put $\tilde{\mu}_1 = \hat{\mu}_{1A}^{(3)}$,

(v) Sometimes-pool prediction decision rule: put $\tilde{\mu}_1 = \hat{\mu}_{1s}$, where $\hat{\mu}_{1s}$ depends on a pooling decision rule,

where $\hat{\mu}_{1N}$ and $\hat{\mu}_{1A}^{(j)}$ are the same as $\hat{p}_{1N}$ and $\hat{p}_{1A}^{(j)}$ in (2.1) except that we replace $\hat{p}_i$ by $\hat{\mu}_i$.

Then we can find a necessary and sufficient condition for which $R(\hat{\mu}_{1A}^{(i)}) < \min[R(\hat{\mu}_{1N}), R(\hat{\mu}_{1A}^{(j)}), R(\hat{\mu}_{1A}^{(k)})]$, $i = 1,2,3$. Based on this necessary and sufficient condition, we can define a sometimes-pool prediction decision rule based on expected loss as we do in Definition 2.1. Then we can show that in the case of equal losses (when $a = b$) the never-pool prediction decision rule and the sometimes-pool prediction decision rule based on expected losses always make the same prediction decision.

CHAPTER 6

A DECISION THEORETIC APPROACH TO THE PROBLEM OF CHOOSING A REGRESSION PREDICTION MODEL

6.1. Introduction.

In Chapter 4, we have shown that the sometimes-pool treatment decision rule based on expected loss and the never-pool treatment decision rule always make the same treatment decision. In Chapter 2, we have mentioned the problem of choosing a regression prediction model. We have mentioned that the deleted-model predictor is sometimes better than the full-model predictor when using the mean square error criterion. We have also mentioned a conditional predictor based on a preliminary test.

In this chapter we discuss the problem of choosing a regression prediction model from another point of view. Instead of predicting the value of a future dependent variable, we are interested in predicting if the future dependent variable is above or below a certain critical number. We thus discretize the problem and are able to discuss the problem in a similar fashion as in the case of pooling data. We show that in the case of equal losses the full-model prediction decision rule and the sometimes deleted-model prediction decision rule based on expected loss always make the same prediction decision. Therefore, the deletion of independent variables is irrelevant in this sense. In the case of unequal losses we give a necessary and sufficient condition for the sometimes deleted-model prediction decision rule to be "better" than the full-model prediction decision rule. Based on this necessary and sufficient condition, we define

a new deleting decision rule. However, this deleting decision rule is irrelevant in the sense that the sometimes deleted-model prediction decision rule based on this deleting decision rule always makes the same prediction decision as the original one does.

In Chapter 2, we have shown that the problem of pooling data in the case of two normal populations with equal variances is a special case of the problem of choosing a regression prediction model when using the mean square error criterion. Again this relation holds in the present decision theoretic framework. In the case of equal variances prediction decisions in the two normal populations case (which is discussed in Section 4.6 for the case of arbitrary unequal variances) can be considered as a special case of the problem of choosing a regression prediction model discussed in this chapter.

6.2. The problem of choosing a regression prediction model.

Suppose we have the following standard linear regression model:

$$(2.1) \qquad \underset{\sim}{Y} = X_1 \underset{\sim}{B}_1 + X_2 \underset{\sim}{B}_2 + \underset{\sim}{e} ,$$

where X_1 and X_2 are known, $E(\underset{\sim}{e}) = \underset{\sim}{0}$ and $Var(\underset{\sim}{e}) = \sigma^2 I$. Suppose we have a future observation Y_0 such that

$$(2.2) \qquad Y_0 = X'_{10} \underset{\sim}{B}_1 + X'_{20} \underset{\sim}{B}_2 + e_0 ,$$

where $\underset{\sim}{X}'_{i0}$, $i = 1,2$, are known, $E(e_0)$ $(e_0) = \sigma^2$, e_0 and $\underset{\sim}{e}$ are independent.

Suppose we want to predict if $Y_0 \leq c$ or $Y_0 > c$ (where c is a real number). Let $\hat{Y}$ be a predictor of Y_0, which is independent of e_0. We can define a prediction decision rule as follows:

(2.3)
(i) Predict that $Y_0 \leq c$ if $\hat{Y} \leq c$
(ii) Predict that $Y_0 > c$ if $\hat{Y} > c$.

Then we are subject to two kinds of error as follows:

(i) Predict that $Y_0 \leq c$ when in fact $Y_0 > c$
(ii) Predict that $Y_0 > c$ when in fact $Y_0 \leq c$.

Hence we have the following loss table.

	$Y_0 \leq c$	$Y_0 > c$
$\hat{Y} \leq c$	0	$a > 0$
$\hat{Y} > c$	$b > 0$	0

Table 2.1

Let

(2.4)
(i) $\hat{Y}_0$ be the full-model predictor as indicated in (3.6) in Chapter 2
(ii) $\hat{\hat{Y}}_0$ be the deleted-model predictor as indicated in (3.7) in Chapter 2.

We define three prediction decision rules as follows:

(2.5)
(i) Full-model prediction decision rule: put $\hat{Y} = \hat{Y}_0$.
(ii) Deleted-model prediction decision rule: put $\hat{Y} = \hat{\hat{Y}}_0$.
(iii) Sometimes deleted-model prediction decision rule: put $\hat{Y} = \hat{Y}^{(s)}$, where $\hat{Y}^{(s)}$ depends on the deleting decision rule which will be discussed later.

Lemma 2.1. The expected loss using $\hat{Y}$ is

$$R(\hat{Y}) = bP(Y_0 \le c) + [a-(a+b)P(Y_0 \le c)]P(\hat{Y} \le c) .$$

Proof: Similar to the proof of Lemma 6.1 in Chapter 4.

Let

$$\mu_0 = \underset{\sim}{X}_{10}'\underset{\sim}{B}_1 + \underset{\sim}{X}_{20}'\underset{\sim}{B}_2, \quad \mu_d = \underset{\sim}{X}_{10}'\underset{\sim}{B}_1 + \underset{\sim}{X}_{10}'(X_1'X_1)^{-1}X_1'X_2\underset{\sim}{B}_2 ,$$

$$\sigma_f^2 = \sigma^2 \underset{\sim}{X}_0'(X'X)^{-1}\underset{\sim}{X}_0, \quad \text{where } \underset{\sim}{X}_0' = [\underset{\sim}{X}_{10}' | \underset{\sim}{X}_{20}'] \text{ and } X = [X_1 | X_2] , \tag{2.6}$$

$$\sigma_d^2 = \sigma^2 \underset{\sim}{X}_{10}'(X_1'X_1)^{-1}\underset{\sim}{X}_{10} .$$

It is easy to see that

$$Y_0 \sim N(\mu_0, \sigma^2), \ \hat{Y}_0 \sim N(\mu_0, \sigma_f^2) \text{ and } \hat{\hat{Y}}_0 \sim N(\mu_d, \sigma_d^2) . \tag{2.7}$$

Theorem 2.1.

$$R(\hat{\hat{Y}}_0) < R(\hat{Y}_0) \Leftrightarrow (B_1, B_2, \sigma^2) \in A_1 \cup A_2 , \tag{2.8}$$

where $A_1 = \{\mu_0 < c - \sigma\gamma \text{ and } \sigma_d\mu_0 - \sigma_f\mu_d > c(\sigma_d - \sigma_f)\}$,

$A_2 = \{\mu_0 > c - \sigma\gamma \text{ and } \sigma_d\mu_0 - \sigma_f\mu_d < c(\sigma_d - \sigma_f)\}$ and $\gamma = \Phi^{-1}[a/(a+b)]$.

Proof: Similar to the proof of Theorem 6.1 in Chapter 4.

Corollary 2.1. If $a = b$ and $\underset{\sim}{B}_2 = \underset{\sim}{0}$, then $R(\hat{\hat{Y}}_0) < R(\hat{Y}_0)$.

Proof: We note that (i) $\gamma = \Phi^{-1}[a/(a+b)] = 0$ when $a = b$,

(ii) $\mu_0 = \mu_d$ when $\underset{\sim}{B}_2 = \underset{\sim}{0}$, (iii) $\sigma_d^2 < \sigma_f^2$ (see Lemma 3.4 in Chapter 2

The result follows from Theorem 2.1. Q.E.D.

We can define a sometimes deleted-model predic[illegible] decision rule based on expected loss as follows:

Definition 2.1. We define a sometimes deleted-model prediction decision rule based on expected loss as follows:

$$(2.9)\quad \begin{cases} \text{Predict that } Y_0 \le c & \text{if } \hat{Y}^{(s)} \le c \\ \text{Predict that } Y_0 > c & \text{if } \hat{Y}^{(s)} > c \ , \end{cases}$$

where the deleting decision rule is as follows:

$$(2.10)\quad \hat{Y}^{(s)} = \begin{cases} \hat{\hat{Y}}_0, & \text{if } (\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in \hat{A}_1 \cup \hat{A}_2 \\ \hat{Y}_0, & \text{otherwise} \ , \end{cases}$$

where

$$(2.11)\quad \begin{aligned} \hat{A}_1 &= [\hat{\mu}_0 < c - \hat{\sigma}\gamma \text{ and } \hat{\sigma}_d\hat{\mu}_0 - \hat{\sigma}_f\hat{\mu}_d > c(\hat{\sigma}_d - \hat{\sigma}_f)] \\ \hat{A}_2 &= [\hat{\mu}_0 > c - \hat{\sigma}\gamma \text{ and } \hat{\sigma}_d\hat{\mu}_0 - \hat{\sigma}_f\hat{\mu}_d < c(\hat{\sigma}_d - \hat{\sigma}_f)] \ , \end{aligned}$$

where $\hat{\mu}_0$, $\hat{\mu}_d$, $\hat{\sigma}_f$ and $\hat{\sigma}_d$ are obtained from μ_0, μ_d, σ_f and σ_d in (2.6) by replacing (B_i, σ^2) by $(\hat{B}_i, \hat{\sigma}^2)$, where $\hat{B}_i$ is the least square estimator of B_i and $\hat{\sigma}^2$ is the standard estimator of σ^2.

Remark: $\hat{A}_i$ is the same as A_i in (2.8) except that we replace (B_1, B_2, σ^2) by $(\hat{B}_1, \hat{B}_2, \hat{\sigma}^2)$.

Lemma 2.2. The full-model prediction decision rule and the deleted-model prediction decision rule have different prediction decisions if and only if

$$(2.12)\quad \begin{aligned} &\text{either (i)} \quad \hat{Y} \le c < \hat{\hat{Y}}_0 \\ &\text{or} \quad\ \text{(ii)} \quad \hat{\hat{Y}}_0 \le c < \hat{Y}_0 \ . \end{aligned}$$

Proof: It follows directly from (2.3) and (2.5).

Lemma 2.3.

$$(2.13)\qquad \begin{array}{ll} \text{(i)} & \hat{\mu}_0 = \hat{Y}_0 \\ \text{(ii)} & \hat{\mu}_d = \hat{\hat{Y}}_0 \ . \end{array}$$

Proof: (i) is clear. (ii) follows from (3.11) in Chapter 2 and some straightforward algebra. Q.E.D.

Lemma 2.4. The full-model prediction decision rule and the deleted-model prediction decision rule have different prediction decisions if and only if

$$(2.14)\qquad (\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in \hat{E}_1 \cup \hat{E}_2 \ ,$$

where $\hat{E}_1 = \{\hat{\mu}_0 \le c < \hat{\mu}_d\}$ and $\hat{E}_2 = \{\hat{\mu}_d \le c < \hat{\mu}_0\}$.

Proof: (2.14) follows directly from (2.12) and (2.13).

Theorem 2.2. Suppose $a = b$. Then the full-model prediction decision rule and the sometimes deleted-model prediction decision rule (defined in Definition 2.1) always make the same prediction.

Proof: Similar to the proof of Theorem 6.2 in Chapter 4.

Corollary 2.2. Suppose $a = b$. Then $R(\hat{Y}_0) = R(\hat{Y}^{(s)})$.

Proof: It follows directly from Theorem 2.2.

So far we have assumed that $a = b$. Now we are going to discuss the case that $a \neq b$.

Lemma 2.5. Assume that $a > b$. Then

$$(2.15)\qquad \begin{array}{ll} \text{(i)} & \hat{A}_2 \ \text{ in (2.11)} = A^* \cup \hat{A}_{21} \cup \hat{A}_{22} \\ \text{(ii)} & \hat{E}_1 \ \text{ in (2.14)} = A^* \cup \hat{E}_{11} \ , \end{array}$$

where

$$A^* = \{c - \hat{\sigma}\gamma < \hat{\mu}_0 \le c \text{ and } \hat{\mu}_d > c\}$$

$$\hat{A}_{21} = \{c - \hat{\sigma}\gamma < \hat{\mu}_0 \le c \text{ and } \hat{\sigma}_d(\hat{\mu}_0 - c)/\hat{\sigma}_f + c < \hat{\mu}_d \le c\}$$

$$\hat{A}_{22} = \{c < \hat{\mu}_0 \text{ and } \hat{\sigma}_d\hat{\mu}_0 - \hat{\sigma}_f\hat{\mu}_d < c(\hat{\sigma}_d - \hat{\sigma}_f)\}$$

$$\hat{E}_{11} = \{\hat{\mu}_0 \le c - \hat{\sigma}\gamma \text{ and } \hat{\mu}_d > c\} .$$

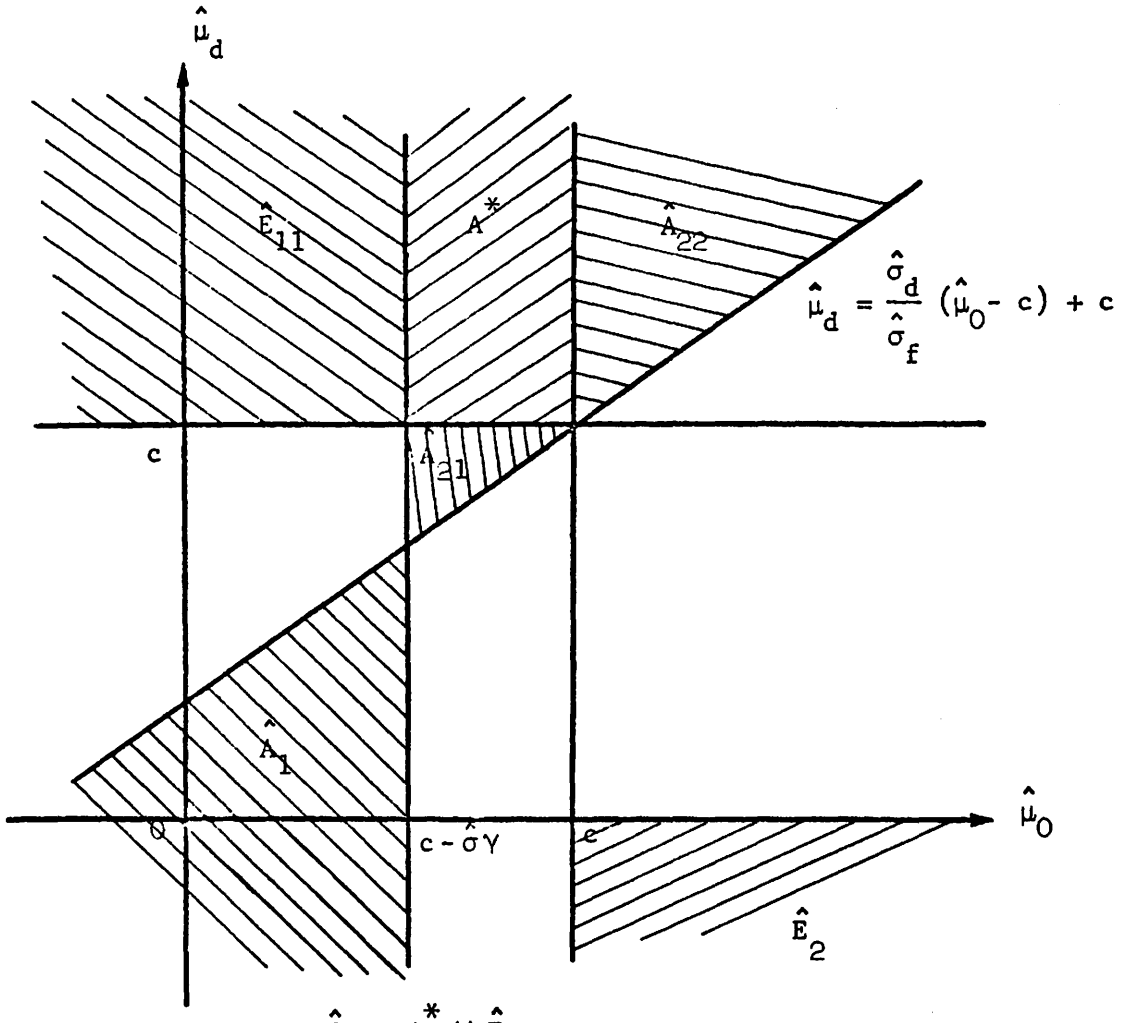
Before we prove Lemma 2.5, let us consider Figure 2.1 (assuming $a > b$) which can help us visualize (2.15) and understand the arguments below. From (2.10), we know that $\hat{Y}^{(s)} = \hat{Y}_0$ if and only if $(\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in \hat{A}_1 \cup \hat{A}_2$. From (2.14), we know that $\hat{Y}_0$ and $\hat{\hat{Y}}_0$ make different prediction decisions if and only if $(\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in \hat{E}_1 \cup \hat{E}_2$. Figure 2.1 indicates that A^* is the intersection of $\hat{A}_1 \cup \hat{A}_2$ with $\hat{E}_1 \cup \hat{E}_2$. It is easy to see that $\hat{Y}_0$ and $\hat{Y}^{(s)}$ make different predictions if and only if $(\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in A^*$. Similar result holds for the case that $a < b$. Lemmas 2.6 and 2.7 and Theorem 2.3 are to state this fact.

Proof of Lemma 2.5: Note that $\gamma = \Phi^{-1}[a/(a+b)] > 0$ when $a > b$. To prove (i), consider $\hat{A}_2 = A^{**} \cup \hat{A}_{22}$, where $A^{**} = \{c - \hat{\sigma}\gamma < \hat{\mu}_0 \le c$ and $\hat{\sigma}_d\hat{\mu}_0 - \hat{\sigma}_f\hat{\mu}_d < c(\hat{\sigma}_d - \hat{\sigma}_f)\} = \{c - \hat{\sigma}\gamma < \hat{\mu}_0 \le c$ and $\hat{\sigma}_d(\hat{\mu}_0 - c)/\hat{\sigma}_f + c < \hat{\mu}_d\}$ $= A^* \cup \hat{A}_{21}$. (ii) is evident. Q.E.D.

Let

$$\mathcal{X} = \{\hat{B}_1, \hat{B}_2, \hat{\sigma}^2 \mid -\infty < \hat{B}_i < \infty \text{ and } \hat{\sigma}^2 > 0\} .$$

Lemma 2.6. Assume that $a > b$. Then the full-model prediction decision rule and the sometimes deleted-model prediction decision rule (defined in Definition 2.1) make the same prediction decisions



$$\hat{E}_1 = A^* \cup \hat{E}_{11}$$

$$\hat{A}_2 = A^* \cup \hat{A}_{21} \cup \hat{A}_{22}$$

Figure 2.1.

if $(\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in \mathcal{Y} - A^*$.

Proof: Assume $(\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in \mathcal{Y} - A^*$. Let $\hat{A} = \hat{A}_1 \cup \hat{A}_{21} \cup \hat{A}_{22}$ and $\hat{E} = \hat{E}_{11} \cup \hat{E}_2$. From (2.10) and (2.15), $\hat{Y}^{(s)} = \hat{\hat{Y}}_0$ iff $(\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in \hat{A}$. From (2.14) and (2.15), $\hat{Y}_0$ and $\hat{\hat{Y}}_0$ make different prediction decisions iff $(\hat{B}_i, \hat{\sigma}^2) \in \hat{E}$. It is sufficient to show that $\hat{A} \subset \hat{E}^c$. This is equivalent to showing that $\hat{E} \subset \hat{A}^c$. Suppose that $(\hat{B}_i, \hat{\sigma}^2) \in \hat{E}_{11}$. Then $\hat{\mu}_0 \leq c - \hat{\sigma}\gamma$ and $\hat{\mu}_d > c$. (i) If $\hat{\mu}_0 = c - \hat{\sigma}\gamma$

and $\hat{\mu}_d > c$, clearly $(\hat{B}_i, \hat{\sigma}^2) \in \hat{A}^c$. (ii) If $\hat{\mu}_0 < c - \hat{\sigma}\gamma$ and $\hat{\mu}_d > c$, then $\hat{\mu}_0 - c < 0 < \hat{\mu}_d - c$. It follows that $\hat{\sigma}_d(\hat{\mu}_0 - c) < \sigma_f(\hat{\mu}_d - c)$. That is $\hat{\sigma}_d\hat{\mu}_0 - \hat{\sigma}_f\hat{\mu}_d < c(\hat{\sigma}_d - \hat{\sigma}_f)$. Hence $(\hat{B}_i, \hat{\sigma}^2) \in \hat{A}^c$. We have shown that $\hat{E}_{11} \subset \hat{A}^c$. Similarly, we can show that $\hat{E}_2 \subset \hat{A}^c$. Q.E.D.

Lemma 2.7. Assume that $a > b$. Then the full-model prediction decision rule and the sometimes deleted-model prediction decision rule (defined in Definition 2.1) make different prediction decisions if $(\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in A^*$.

Proof: By (2.10) and (2.15), $\hat{Y}^{(s)} = \hat{\hat{Y}}_0$ if $(\hat{B}_i, \hat{\sigma}^2) \in A^*$. By (2.14) and (2.15), $\hat{Y}_0$ and $\hat{\hat{Y}}_0$ make different prediction decisions if $(\hat{B}_i, \hat{\sigma}^2) \in A^*$. The result follows. Q.E.D.

Theorem 2.3. Assume that $a \neq b$. Then the full-model prediction decision rule and the sometimes deleted-model prediction decision rule (defined in Definition 2.1) make different prediction decisions if and only if

$$(2.16) \qquad (\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in K^* ,$$

where

$$(2.17) \qquad K^* = \begin{cases} A^*, & \text{if } a > b \\ (c < \hat{\mu}_0 < c - \hat{\sigma}\gamma \text{ and } \hat{\mu}_d \leq c), & \text{if } a < b \end{cases}$$

$$\gamma = \Phi^{-1}[a/(a+b)] .$$

Proof: If $a > b$, the result follows directly from Lemmas 2.6 and 2.7. Similar arguments for the case $a < b$. Q.E.D.

Theorem 2.4.

(i) If $a > b$, then $R(\hat{Y}^{(s)}) < R(\hat{Y}_0) \Leftrightarrow \mu_0 > c - \sigma\gamma$.

(ii) If $a < b$, then $R(\hat{Y}^{(s)}) < R(\hat{Y}_0) \Leftrightarrow \mu_0 < c - \sigma\gamma$.

Proof: To prove (i), consider that $K^* = A^* = \{c-\hat{\sigma}\gamma < \hat{\mu}_0 \leq c$ and $\hat{\mu}_d > c\}$. Let $R(\cdot) = R_{K^*}(\cdot) + R_{\mathcal{Y}-K^*}(\cdot)$, where the terms on the right correspond to integrals over K^* and $\mathcal{Y} - K^*$. By Theorem 2.3, $R_{\mathcal{Y}-K^*}(\hat{Y}^{(s)}) = R_{\mathcal{Y}-K^*}(\hat{Y}_0)$. By (2.10), (2.13), (2.15) and (2.17), $(\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in K^* \Rightarrow$

(i) $\hat{Y}^{(s)} = \hat{\hat{Y}}_0$ and $\hat{\hat{Y}}_0 > c$

(ii) $\hat{Y}_0 \leq c$.

Therefore, when $(\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in K^*$, we have the following losses:

(i)

$Y_0 \leq c$	$Y_0 > c$
b	0

when using $\hat{Y}^{(s)}$

(ii)

$Y_0 \leq c$	$Y_0 > c$
0	a

when using $\hat{Y}_0$.

Conseuquently, $R_{K^*}(\hat{Y}^{(s)}) = bP(Y_0 \leq c \text{ and } K^*) = bP(Y_0 \leq c)P(K^*)$ and $R_{K^*}(\hat{Y}_0) = aP(Y_0 > c \text{ and } K^*) = a[1-P(Y_0 \leq c)]P(K^*)$. Now it is easy to see that $R(\hat{Y}^{(s)}) < R(\hat{Y}_0) \Leftrightarrow R_{K^*}(\hat{Y}^{(s)}) < R_{K^*}(\hat{Y}_0) \Leftrightarrow \mu_0 > c-\sigma\gamma$. We have proved (i). Similar proof gives (ii). Q.E.D.

When $a \neq b$, Theorem 2.4 tells us that the sometimes deleted-model prediction decision rule (defined in Definition 2.1) is sometimes better than the full-model prediction decision rule, but sometimes not. Can we find a better sometimes deleted-model prediction decision rule? We may define another sometimes deleted-

model prediction decision rule based on Theorem 2.4 as follows.

Definition 2.2. Assume $a \neq b$. We define a sometimes deleted-model prediction decision rule (based on Theorem 2.4) as follows:

$$(2.18)\quad \begin{cases} \text{Predict that } Y_0 \leq c \text{ if } \hat{\hat{Y}}^{(s)} < c \\ \text{Predict that } Y_0 > c \text{ if } \hat{\hat{Y}}^{(s)} > c \ , \end{cases}$$

where the deleting decision rule is as follows:

$$(2.19)\quad \hat{\hat{Y}}^{(s)} = \begin{cases} \hat{Y}^{(s)}, & \text{if } (\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in \hat{\hat{A}} \\ \hat{Y}_0, & \text{otherwise} \ , \end{cases}$$

where

$$(2.20)\quad \hat{\hat{A}} = \begin{cases} (\hat{\mu}_0 > c - \hat{\sigma}\gamma) & \text{when } a > b \\ (\hat{\mu}_0 < c - \hat{\sigma}\gamma) & \text{when } a < b \ . \end{cases}$$

From (2.10), (2.11), (2.19) and (2.20), it follows that $\hat{\hat{Y}}^{(s)}$ in (2.19) is equivalent to the following:

$$(2.21)\quad \hat{\hat{Y}}^{(s)} = \begin{cases} \hat{\hat{Y}}_0, & \text{if } (\hat{B}_1, \hat{B}_2, \hat{\sigma}^2) \in \begin{cases} \hat{A}_2 & \text{when } a > b \\ \hat{A}_1 & \text{when } a < b \end{cases} \\ \hat{Y}_0, & \text{otherwise} \ . \end{cases}$$

Then we have the following result.

Theorem 2.5. Assume $a \neq b$. $\hat{Y}^{(s)}$ in (2.10) and $\hat{\hat{Y}}^{(s)}$ in (2.19) always make the same prediction decision.

Proof: Suppose $a > b$. By (2.10) and (2.21), it is sufficient to show that $\hat{Y}_0$ and $\hat{\hat{Y}}_0$ make the same prediction decision if $(\hat{B}_i, \hat{\sigma}^2) \in \hat{A}_1$. Theorem 2.3 tells us that $\hat{Y}_0$ and $\hat{\hat{Y}}_0$ make the same prediction decision iff $(\hat{B}_i, \hat{\sigma}^2) \in \Upsilon - A^*$. By (2.15),

$A^* \subset \hat{A}_2$. By (2.11), we note that $\hat{A}_1$ and $\hat{A}_2$ are disjoint. It follows that $\hat{A}_1 \subset \mathcal{I} - A^*$. Similar proof for the case $a < b$. Q.E.D.

Corollary 2.3. Assume $a \neq b$. Then $R(\hat{Y}^{(s)}) = R(\hat{\hat{Y}}^{(s)})$, where $\hat{Y}^{(s)}$ and $\hat{\hat{Y}}^{(s)}$ are respectively defined in (2.10) and (2.19).

Proof: It follows directly from Theorem 2.5.

Remark: Theorem 2.5 tells us that our new deleting decision rule (2.19) based on Theorem 2.4 is irrelevant. The new sometimes-deleted model prediction decision rule (defined in Definition 2.2) always makes the same prediction decision as the old one (defined in Definition 2.1) does. In other words, we don't find another sometimes deleted-model prediction decision rule which is better than the one defined in Definition 2.1.

CHAPTER 7

A BAYESIAN APPROACH TO THE PROBLEM OF POOLING DATA AND THE PROBLEM OF CHOOSING A REGRESSION PREDICTION MODEL

7.1. Introduction.

Suppose that there are two populations having parameters θ_1, θ_2. We are interested in estimating θ_1. Suppose we have past data available from both populations. Let $\hat{\theta}_{1N}$ be an estimator based on the observations from the first population only. $\hat{\theta}_{1N}$ is called the never-pooled estimator. Let $\hat{\theta}_{1A}$ be an estimator based on the observations from both populations. $\hat{\theta}_{1A}$ is called the always-pooled estimator. It is well known that $E(\hat{\theta}_{1A} - \theta_1)^2$ can be less than $E(\hat{\theta}_{1N} - \theta_1)^2$ for some parameter values.

Suppose that we have some prior knowledge about the values of θ_1 and θ_2 and the knowledge can be expressed in probabilistic form, say $\pi(\theta_1, \theta_2)$. We may have a loss function, say $L(\theta_1, \tilde{\theta}_1)$, where $\tilde{\theta}_1$ is an estimator of θ_1 depending on the given sample $\underset{\sim}{Y} = (Y_1, Y_2, \ldots, Y_n)$. $\underset{\sim}{Y}$ are observations either from the first population only or from both populations. The risk function associated with the estimator $\tilde{\theta}_1$ is given by

$$\gamma(\theta_1, \theta_2) = \int_R L(\theta_1, \tilde{\theta}_1) p(\underset{\sim}{Y} | \theta_1, \theta_2) d\underset{\sim}{Y},$$

where $p(\underset{\sim}{Y} | \theta_1, \theta_2)$ is a proper p.d.f. for $\underset{\sim}{Y}$ given (θ_1, θ_2) and R is the range of $\underset{\sim}{Y}$. The Bayes risk associated with the estimator $\tilde{\theta}_1$ is defined by:

$$E\gamma(\theta_1, \theta_2) = \iint \gamma(\theta_1, \theta_2)\pi(\theta_1, \theta_2)d\theta_1 d\theta_2 .$$

Suppose that we have a standard regression model, $\underset{\sim}{Y} = X_1\underset{\sim}{B}_1 + X_2\underset{\sim}{B}_2 + \underset{\sim}{e}$. Suppose that we have a future observation, say $Y_0 = \underset{\sim}{X}_{10}' \underset{\sim}{B}_1 + \underset{\sim}{X}_{20}' \underset{\sim}{B}_2 + e_0$. As we have mentioned in Chapter 2, we can use either the full-model predictor, $\hat{Y}_0 = \underset{\sim}{X}_{10}'\underset{\sim}{\hat{B}}_1 + \underset{\sim}{X}_{20}'\underset{\sim}{\hat{B}}_2$, or the deleted-model predictor, $\hat{\hat{Y}}_0 = \underset{\sim}{X}_{10}' \underset{\sim}{\hat{\hat{B}}}_1$, to predict Y_0. It is well known that $E(\hat{\hat{Y}}_0 - Y_0)^2$ can be less than $E(\hat{Y}_0 - Y_0)^2$ for some parameter values.

Suppose that we have some prior knowledge about $(\underset{\sim}{B}_1, \underset{\sim}{B}_2)$ and this knowledge can be expressed in probabilistic form, say $\pi(\underset{\sim}{B}_1, \underset{\sim}{B}_2)$. We also have a loss function $L(Y_0, \tilde{Y}_0)$, where $\tilde{Y}_0 = \underset{\sim}{X}_{10}'\underset{\sim}{\tilde{B}}_1 + \underset{\sim}{X}_{20}'\underset{\sim}{\tilde{B}}_2$, and $(\underset{\sim}{\tilde{B}}_1, \underset{\sim}{\tilde{B}}_2)$ depends on $\underset{\sim}{Y}$. Similarly, the risk function associated with the predictor $\tilde{Y}_0$ is given by

$$\gamma(\underset{\sim}{B}_1, \underset{\sim}{B}_2) = \int_R L(Y_0, \tilde{Y}_0)p(\underset{\sim}{Y}|\underset{\sim}{B}_1, \underset{\sim}{B}_2)d\underset{\sim}{Y} ,$$

where $p(\underset{\sim}{Y}|\underset{\sim}{B}_1, \underset{\sim}{B}_2)$ is a proper p.d.f. for $\underset{\sim}{Y}$ given $(\underset{\sim}{B}_1, \underset{\sim}{B}_2)$ and R is the range of $\underset{\sim}{Y}$. The Bayes risk associated with the predictor $\tilde{Y}_0$ is:

$$E\gamma(\underset{\sim}{B}_1, \underset{\sim}{B}_2) = \iint \gamma(\underset{\sim}{B}_1, \underset{\sim}{B}_2)\pi(\underset{\sim}{B}_1, \underset{\sim}{B}_2)d\underset{\sim}{B}_1 d\underset{\sim}{B}_2 .$$

In our discussions, we always assume the quadratic loss, i.e. $L(\tilde{\theta}_1, \theta_1) = (\tilde{\theta}_1 - \theta_1)^2$ and $L(Y_0, \tilde{Y}_0) = (Y_0 - \tilde{Y}_0)^2$.

In Section 7.2, we consider the case of two normal populations. We derive a necessary and sufficient condition for which the Bayes risk associated with the always-pooled estimator is less than the Bayes risk associated with the never-pooled estimator. Assuming

normal prior, we derive the Bayes estimator which minimizes the Bayes risk. In Section 7.3, we derive a necessary and sufficient condition for which the Bayes risk associated with the deleted-model predictor is less than the Bayes risk associated with the full-model predictor. Also assuming normal prior, we derive the Bayes predictor which minimizes the Bayes risk. In Section 7.4, we show that for two normal populations the Bayes risk of the deleted-model predictor is less than that of the full-model predictor if and only if the Bayes risk of the always-pooled estimator is less than that of the never-pooled estimator. The always-pooled, the never-pooled and the Bayes estimators correspond respectively to the deleted-model, the full-model and the Bayes predictors. In Sections 7.5 and 7.6, we assume independent priors in the binomial and the Poisson cases. A linear combination of the never-pooled and the always-pooled estimators is "better" than the never-pooled, but not the "best" or "Bayes" solution. The Bayes estimator does not involve observations from the second population.

The proofs of propositions in this chapter are routine and straightforward. The proofs are given in Appendix D.

7.2. The problem of pooling data for two normal populations.

Suppose we have two normal populations, say $N_i(\mu_i, \sigma^2)$, $i = 1,2$. We are interested in estimating μ_1. Suppose we have the following past data available: $(Y_{ij}, j = 1,2,\ldots,n_i)$ from $N_i(\mu_i, \sigma^2)$, $i = 1,2$. Y_{ij}'s are independent. Let $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$,

$i = 1,2$. We have two estimators for μ_1 as follows:

(i) Never-pooled estimator $\hat{\mu}_{1N} = \bar{Y}_1$

(ii) Always-pooled estimator $\hat{\mu}_{1A} = (n_1\bar{Y}_1 + n_2\bar{Y}_2)/(n_1 + n_2)$.

Suppose that the prior distribution of (μ_1, μ_2) is an arbitrary distribution with

$$(2.1) \qquad E\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix} = \begin{bmatrix}\overset{0}{\mu_1} \\ \overset{0}{\mu_2}\end{bmatrix} \text{ and } \operatorname{Var}\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix} = \begin{bmatrix}\overset{0}{\sigma_1^2} & \overset{0}{\sigma_{12}} \\ \overset{0}{\sigma_{12}} & \overset{0}{\sigma_2^2}\end{bmatrix} .$$

The following proposition gives us a necessary and sufficient condition for which the Bayes risk associated with $\hat{\mu}_{1A}$ is less than the Bayes risk associated with $\hat{\mu}_{1N}$.

Proposition 2.1. Assume that (i) the prior distribution of (μ_1, μ_2) is as indicated above, (ii) σ^2 is known. Then

$$(2.2) \qquad \begin{aligned} &EE[(\hat{\mu}_{1A} - \mu_1)^2 | \mu_1, \mu_2] < EE[(\hat{\mu}_{1N} - \mu_1)^2 | \mu_1, \mu_2] \Leftrightarrow \\ &[n_1 n_2/(n_1 + n_2)][(\overset{0}{\mu_1} - \overset{0}{\mu_2})^2 + (\overset{0}{\sigma_1^2} + \overset{0}{\sigma_2^2} - 2\overset{0}{\sigma_{12}})] < \sigma^2 . \end{aligned}$$

The next proposition shows that the Bayes estimator of μ_1 is a linear combination of $\bar{Y}_1$ and $\bar{Y}_2$.

Proposition 2.2. Assume that (i) the prior distribution of (μ_1, μ_2) is a bivariate normal distribution with mean and positive-definite covariance matrix as indicated in (2.1), (ii) σ^2 is known. Let $\hat{\mu}_1 = \hat{\mu}_1(\bar{Y}_1, \bar{Y}_2)$, a function of $(\bar{Y}_1, \bar{Y}_2)$. Then

$$(2.3) \qquad \min_{\hat{\mu}_1} EE[(\hat{\mu}_1 - \mu_1)^2 | \mu_1, \mu_2] = EE[(\hat{\mu}_1^* - \mu_1)^2 | \mu_1, \mu_2] ,$$

where

$$\hat{\mu}_1^* = a + b\bar{Y}_1 + c\bar{Y}_2 ,$$

$$a = (1/k)[(\sigma^4 + n_2\sigma_2^{0^2}\sigma^2)\mu_1^0 - n_2\sigma^2\sigma_{12}^0\mu_2^0] ,$$

(2.4) $$b = (1/k)[n_1\sigma^2\sigma_1^{0^2} + n_2n_1(\sigma_1^{0^2}\sigma_2^{0^2} - \sigma_{12}^{0^2})],$$

$$c = (1/k)[n_2\sigma^2\sigma_{12}^0] ,$$

$$k = \sigma^4 + \sigma^2(n_1\sigma_1^{0^2} + n_2\sigma_2^{0^2}) + n_1n_2(\sigma_1^{0^2}\sigma_2^{0^2} - \sigma_{12}^{0^2}) .$$

Corollary 2.1. Suppose that the assumptions in Proposition 2.2 hold except that $\sigma_{12}^0 = 0$. Then

(2.5) $$\hat{\mu}_1^* = a' + b'\bar{Y}_1 ,$$

where a' and b' are the same as a and b in (2.4) except that $\sigma_{12}^0 = 0$.

Proof: It follows directly from Proposition 2.2.

Remark: When the prior distribution of (μ_1, μ_2) is uncorrelated, the Bayes estimator of μ_1 does not involve observations from the second population, as we expect.

7.3. The problem of choosing a regression prediction model.

Suppose we have the following standard regression model:

(3.1) $$\underset{\sim}{Y} = X_1\underset{\sim}{B}_1 + X_2\underset{\sim}{B}_2 + \underset{\sim}{e} ,$$

where X_1 and X_2 are known, $E(\underset{\sim}{e}) = 0$ and $Var(\underset{\sim}{e}) = \sigma^2 I$. Suppose we have a future observation Y_0 such that

(3.2) $$Y_0 = \underset{\sim}{X}_{10}'\underset{\sim}{B}_1 + \underset{\sim}{X}_{20}'\underset{\sim}{B}_2 + e_0 ,$$

where $\underset{\sim}{X}_{i0}$, $i = 1,2$, are known, $E(e_0) = 0$ and $Var(e_0) = \sigma^2$.

e_0 is independent of $\underset{\sim}{e}$. We want to predict Y_0. As we have mentioned in Chapter 2, we have two predictors as follows:

(i) The full-model predictor, $\hat{Y}_0$, as indicated in (3.6) (3.3) in Chapter 2

(ii) The deleted-model predictor, $\hat{\hat{Y}}_0$, as indicated in (3.7) in Chapter 2.

Suppose that the prior distribution of $\underset{\sim}{B}' = [\underset{\sim}{B}_1' | \underset{\sim}{B}_2']$ is an arbitrary distribution with

$$(3.4) \quad E\underset{\sim}{B} = \underset{\sim}{B}^0 = \begin{bmatrix} \underset{\sim}{B}_1^0 \\ \underset{\sim}{B}_2^0 \end{bmatrix} \text{ and } \operatorname{Var} \underset{\sim}{B} = \Sigma^0 = \begin{bmatrix} \Sigma_{11}^0 & \Sigma_{12}^0 \\ \Sigma_{12}^0 & \Sigma_{22}^0 \end{bmatrix}.$$

The following proposition gives us a necessary and sufficient condition for which the Bayes risk associated with $\hat{\hat{Y}}_0$ is less than the Bayes risk associated with $\hat{Y}_0$.

Proposition 3.1. Assume that (i) the prior distribution of $\underset{\sim}{B}$ is as indicated above, (ii) σ^2 is known. Then

$$(3.5) \quad \begin{aligned} &EE[(\hat{\hat{Y}}_0 - Y_0)^2 | \underset{\sim}{B}] < EE[(\hat{Y}_0 - Y_0)^2 | \underset{\sim}{B}] \Leftrightarrow \\ &\underset{\sim}{Z}_0' \Sigma_{22}^0 \underset{\sim}{Z}_0 + \underset{\sim}{Z}_0' \underset{\sim}{B}_2^0 \underset{\sim}{B}_2^{0\prime} \underset{\sim}{Z}_0 < \sigma^2 \underset{\sim}{Z}_0' W^{-1} \underset{\sim}{Z}_0 , \end{aligned}$$

where $\underset{\sim}{Z}_0$ and W are the same as indicated in (3.9) and (3.10) in Chapter 2.

Next, we would like to find the Bayes estimator for $\underset{\sim}{B}$. Suppose we have the following linear regression model:

$$\underset{\sim}{Y} = X_1 \underset{\sim}{B}_1 + \underset{\sim}{B}_2 + \underset{\sim}{e} ,$$

with the following assumptions:

(3.6)

(i) $\underset{\sim}{e} \sim N(0, \Delta)$, Δ is known.

(ii) The prior distribution of $\underset{\sim}{B}' = [\underset{\sim}{B}_1' | \underset{\sim}{B}_2']$ is a multivariate normal distribution with mean and positive-definite matrix as indicated in (3.4).

(iii) $\underset{\sim}{B}$ and $\underset{\sim}{e}$ are independent.

The following proposition gives us the Bayes estimator for $\underset{\sim}{B}' = [\underset{\sim}{B}_1' | \underset{\sim}{B}_2']$.

Proposition 3.2. Suppose the assumptions indicated in (3.6) hold. Then out of the class of estimators $\tilde{\underset{\sim}{B}} = f(\underset{\sim}{Y})$, the estimator which minimizes the mean square error $E(\tilde{\underset{\sim}{B}} - \underset{\sim}{B})(\tilde{\underset{\sim}{B}} - \underset{\sim}{B})'$ in the positive semi-definite sense is given by

$$(3.7) \qquad \hat{\underset{\sim}{B}}^* = E[\underset{\sim}{B} | \underset{\sim}{Y}] = \underset{\sim}{B}^0 + \Delta^0 X'(X \Delta^0 X' + \Delta)^{-1}(\underset{\sim}{Y} - X\underset{\sim}{B}^0) .$$

Suppose we have a future observation $Y_0 = \underset{\sim}{X}_{10}'\underset{\sim}{B}_1 + \underset{\sim}{X}_{20}'\underset{\sim}{B}_2 + \underset{\sim}{e}$, where

$$(3.8) \qquad E(e_0) = 0, \quad Var(e_0) = \sigma^2 \quad \text{and} \quad e_0 \text{ is independent of } \underset{\sim}{e} \text{ and } \underset{\sim}{B}'.$$

The following proposition gives us the Bayes predictor for Y_0.

Proposition 3.3. Suppose (3.6) and (3.8) hold. Then out of the class of predictors $\tilde{Y}_0 = \underset{\sim}{X}_0' \tilde{\underset{\sim}{B}}$, where $\underset{\sim}{X}'_0 = [\underset{\sim}{X}_{10}' | \underset{\sim}{X}_{20}']$ and $\tilde{\underset{\sim}{B}} = f(\underset{\sim}{Y})$, the predictor which minimizes the mean square error $E(Y_0 - \tilde{Y}_0)^2$ is given by

$$(3.9) \qquad \hat{Y}_0^* = \underset{\sim}{X}_0' \hat{\underset{\sim}{B}}^* ,$$

where $\hat{\underset{\sim}{B}}^*$ is the same as indicated in (3.7).

Remark: The deleted-model predictor $\hat{\hat{Y}}_0 = \underset{\sim}{X}_{10}' \hat{\underset{\sim}{B}}_1$ is a special case of $\underset{\sim}{X}_0' \tilde{\underset{\sim}{B}}$ if we put $\tilde{\underset{\sim}{B}} = \begin{bmatrix} \hat{\underset{\sim}{B}}_1 \\ 0 \end{bmatrix}$.

7.4. The relationship between the problem of pooling data and the problem of choosing a regression prediction model.

We have shown in Chapter 2 that we can discuss the problem of pooling data in the framework of the problem of choosing a regression prediction model. Again, this relation can be established when we use the Bayesian approach.

Suppose we have the following past data available: n_i observations $(Y_{ij}, j = 1,2,\ldots,n_i)$ from $N_i(\mu_i,\sigma^2)$, $i = 1,2$. Y_{ij}'s are independent. Our main interest is to estimate μ_1. Let

$$Y_j = Y_{1j}, \quad j = 1,2,\ldots,n_1$$
$$Y_{n_1+j} = Y_{2j}, \quad j = 1,2,\ldots,n_2 .$$

Let

$$\alpha = \mu_2, \quad \beta = \mu_1 - \mu_2 . \tag{4.1}$$

Then

$$Y_j = \alpha + \beta X_j + e_j, \quad j = 1,2,\ldots,n_1+ n_2 . \tag{4.2}$$

where

$$X_j = \begin{cases} 1, & j = 1,2,\ldots,n_1 \\ 0, & j = n_1+1,\ldots,n_1+ n_2 \end{cases}$$

e_j's are independently distributed as $N(0,\sigma^2)$.

Suppose we delete the X_j variable. Then

$$Y_j = \alpha + e_j, \quad j = 1,2,\ldots,n_1+ n_2 . \tag{4.3}$$

Let

(4.4) $\hat{\alpha}$ and $\hat{\beta}$ be the least square estimators for α and β in (4.2)
$\hat{\alpha}$ be the least square estimator for α in (4.3).

In Chpater 2 (Lemma 4.1), we have shown that

(4.5) (i) the never-pooled estimator $\hat{\mu}_{1N} = \hat{\alpha} + \hat{\beta} = \sum_{j=1}^{n_1} Y_j/n_1$

(ii) the always-pooled estimator $\hat{\mu}_{1A} = \hat{\hat{\alpha}} = \sum_{j=1}^{n_1+n_2} Y_j/(n_1+n_2)$.

Suppose that the prior distribution of (μ_1, μ_2) is an arbitrary distribution with mean and covariance matrix as indicated in (2.1). From (4.1), it follows that the prior distribution of (α, β) is an arbitrary distribution with

$$(4.6) \qquad E\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = \begin{bmatrix}\mu_2^0\\ \mu_1^0 - \mu_2^0\end{bmatrix} \text{ and } \mathrm{Var}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = \begin{bmatrix}\sigma_2^{0^2} & \sigma_{12}^0 - \sigma_2^{0^2}\\ \sigma_{12}^0 - \sigma_2^{0^2} & \sigma_1^{0^2} + \sigma_2^{0^2} - 2\sigma_{12}^0\end{bmatrix}.$$

Suppose that Y_0 is a future observation such that

$$(4.7) \qquad Y_0 = \alpha + \beta + e_0 ,$$

where e_0 is independent of e_j. Then the full-model and the deleted-model predictors for Y_0 are respectively as follows:

(4.8) (i) $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}$

(ii) $\hat{\hat{Y}}_0 = \hat{\hat{\alpha}}$.

Then we have the following result.

Proposition 4.1. Assume that (i) the prior distribution of (α, β) is an arbitrary distribution with mean and covariance matrix as indicated in (4.6), (ii) σ^2 is known. Then, a necessary and sufficient condition for $EE[(\hat{\hat{Y}}_0 - Y_0)^2|\alpha,\beta] < EE[(\hat{Y}_0 - Y_0)^2|\alpha,\beta]$ is the same as stated in (2.2).

Remark: Lemma 4.1 shows that in a sense, the never-pooled and the always-pooled estimators correspond respectively to the full-model and the deleted-model predictors.

Next, we would like to apply Proposition 3.2 to find the Bayes estimator for μ_1.

Suppose that the prior distribution of (μ_1, μ_2) is a bivariate normal distribution with mean and covariance matrix as indicated in (2.1).

Let $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$, $i = 1,2$. Let $\bar{e}_i = \bar{Y}_i - \mu_i$, $i = 1,2$. Our assumptions imply that conditional on (μ_1, μ_2), (i) $\bar{e}_i \sim N(0, \sigma^2/n_i)$, (ii) $\bar{e}_1$ and $\bar{e}_2$ are independent. It is easy to see that (i) and (ii) still hold unconditional on (μ_1, μ_2). Also (μ_1, μ_2) is independent of $(\bar{e}_1, \bar{e}_2)$.

Let

$$(4.9) \qquad \gamma = \mu_1, \quad \tau = \mu_2 - \mu_1 .$$

Then

$$(4.10) \qquad \bar{Y}_1 = \gamma + \bar{e}_1, \quad \bar{Y}_2 = \gamma + \tau + \bar{e}_2 .$$

Writing (4.10) in matrix form gives the following regression model:

$$(4.11) \qquad \begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \gamma \\ \tau \end{bmatrix} + \begin{bmatrix} \bar{e}_1 \\ \bar{e}_2 \end{bmatrix} .$$

From (4.9), it follows that the prior distribution of (γ,τ) is a bivariate normal distribution with

$$(4.12) \qquad E\begin{bmatrix} \gamma \\ \tau \end{bmatrix} = \begin{bmatrix} \mu_1^0 \\ \mu_2^0 - \mu_1^0 \end{bmatrix} \text{ and } \operatorname{Var}\begin{bmatrix} \gamma \\ \tau \end{bmatrix} = \begin{bmatrix} \sigma_1^{0^2} & \sigma_{12}^0 - \sigma_1^{0^2} \\ \sigma_{12}^0 - \sigma_1^{0^2} & \sigma_1^{0^2} + \sigma_2^{0^2} - 2\sigma_{12}^0 \end{bmatrix} .$$

Then we can apply Proposition 3.2 to find the Bayes estimator for γ in this particular regression model (4.11). We have the following result.

Proposition 4.2. Suppose that we have a regression model as indicated in (4.11). Assume that (i) the prior distribution of (γ, τ) is as indicated above, (ii) σ^2 is known. Let $\hat{\gamma} = \hat{\gamma}(\bar{Y}_1, \bar{Y}_2)$, a function of $(\bar{Y}_1, \bar{Y}_2)$. Then

$$(4.13) \qquad \min_{\hat{\gamma}} EE[(\hat{\gamma}-\gamma)^2|\gamma,\tau] = EE[(\hat{\gamma}^*-\gamma)^2|\gamma,\tau]$$

when

$$(4.14) \qquad \hat{\gamma}^* = a + b\bar{Y}_1 + c\bar{Y}_2 ,$$

where a, b and c are given in (2.4).

Remark: To find the Bayes estimator for μ_1, we have two procedures.

Procedure I: Following the routine work, find the posterior distribution of μ_1. The mean of the posterior distribution is the Bayes estimator (as indicated in Proposition 2.2).

Procedure II: Put it in the framework of linear regression model. Then apply Proposition 3.2 to obtain the Bayes estimator (as indicated in Proposition 4.2). Procedure II involves less algebraic work than Procedure I.

7.5. The problem of pooling data for two binomial populations.

Suppose we have two binomial populations, say π_i with parameter p_i, $i = 1,2$. We are interested in estimating p_1.

We can consider p_i as the probability of failure in each Bernoulli trial. Suppose we have the following past data available:

X_i defectives out of n_i observations from π_i, $i = 1,2$.

X_1 and X_2 are independent.

We have two estimators for p_1 as follows:

(5.1) (i) Never-pool estimator $\hat{p}_{1N} = X_1/n_1$

(ii) Always-pool estimator $\hat{p}_{1A} = (X_1 + X_2)/(n_1 + n_2)$.

The following proposition shows that even under the assumption of independent priors a linear combination of $\hat{p}_{1N}$ and $\hat{p}_{1A}$ is better than $\hat{p}_{1N}$ (but not the "best" or "Bayes" solution).

Proposition 5.1. Assume that p_1 and p_2 are independent a priori and the prior distribution of p_i is Beta (ν_1, ν_2), $i = 1,2$. Let $\hat{p}_{1s} = a\hat{p}_{1N} + (1-a)\hat{p}_{1A}$, $0 \le a \le 1$. Then

(5.2) $$\min_a EE[(\hat{p}_{1s} - p_1)^2 | p_1, p_2] = EE[(\hat{p}^0_{1s} - p_1)^2 | p_1, p_2] ,$$

where

(i) $\hat{p}^0_{1s} = a_0\hat{p}_{1N} + (1-a_0)\hat{p}_{1A}$ (assuming that $\hat{p}_{1N} \neq \hat{p}_{1A}$)

$$a_0 = \begin{cases} 0, & \text{if } a_{00} \le 0 \\ a_{00}, & \text{if } 0 < a_{00} \le 1 \\ 1, & \text{if } a_{00} > 1 \end{cases}$$

$$a_{00} = [(X_1 + \nu_1)/(n_1 + \nu_1 + \nu_2) - \hat{p}_{1A}]/(\hat{p}_{1N} - \hat{p}_{1A}) ,$$

(ii) $\hat{p}^0_{1s} = \hat{p}_{1N} = \hat{p}_{1A}$ (if $\hat{p}_{1N} = \hat{p}_{1A}$) .

The next proposition shows that the Bayes estimator for p_1 does not depend on X_2 under the assumption of independent priors.

Proposition 5.2. Assume that p_1 and p_2 are independent a priori and the prior distribution of p_i is Beta (ν_1,ν_2), $i = 1,2$. Let $\hat{p}_1 = \hat{p}_1(X_1,X_2)$, a function of X_1 and X_2. Then

$$(5.3) \qquad \min_{p_1} EE[(\hat{p}_1-p_1)^2|p_1,p_2] = EE[(\hat{p}_{10}-p_1)^2|p_1,p_2] ,$$

where

$$\hat{p}_{10} = (X_1+\nu_1)/(n_1+\nu_1+\nu_2) .$$

Following Propositions 5.1 and 5.2, we have the following result.

Proposition 5.3. Assume that p_1 and p_2 are independent a priori and the prior distribution of p_i is Beta (ν_1,ν_2), $i = 1,2$. Then

$$(5.4) \qquad EE[(\hat{p}_{10}-p_1)^2|p_1,p_2] < EE[(\hat{p}^0_{1s}-p_1)^2|p_1,p_2] \le EE[(\hat{p}_{1N}-p_1)^2|p_1,p_2],$$

where $\hat{p}_{10}$, $\hat{p}^0_{1s}$ and $\hat{p}_{1N}$ are indicated respectively in (5.3), (5.2) and (5.1).

Corollary 5.1. Assume that p_1 and p_2 are independent a priori and the prior distribution of p_i is uniform on $(0,1)$, $i = 1,2$. Then

$$(5.5) \qquad EE[(\hat{p}_{10}-p_1)^2|p_1,p_2] < EE[(\hat{p}^0_{1s}-p_1)^2|p_1,p_2] \le EE[(\hat{p}_{1N}-p_1)^2|p_1,p_2],$$

where

$\hat{p}_{10} = (X_1+1)/(n_1+2)$,

$\hat{p}^0_{1s}$ is the same as indicated in (5.2) except that

$a_{00} = [(X_1+1)/(n_1+2)-\hat{p}_{1A}]/(\hat{p}_{1N}-\hat{p}_{1A})$,

$\hat{p}_{1N}$ is as indicated in (5.1).

Proof: Note that uniform distribution on $(0,1)$ is Beta $(1,1)$. The result follows immediately from Proposition 5.3.

7.6. The problem of pooling data for two Poisson populations.

Suppose we have two Poisson populations, say $P_i(\lambda_i)$, $i = 1,2$. We are interested in estimating λ_1. Suppose we have the following past data available:

n_i observations $(Y_{ij}, j = 1,2,\ldots,n_i)$ from $P_i(\lambda_i)$, $i = 1,2$;

Y_{ij}'s are independent.

Let $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$. We have two estimators for λ_1 as follows:

(6.1) (i) Never-pooled estimator $\hat{\lambda}_{1N} = \bar{Y}_1$

(ii) Always-pooled estimator $\hat{\lambda}_{1A} = (n_1\bar{Y}_1 + n_2\bar{Y}_2)/(n_1 + n_2)$.

Again, we will show that even under the assumption of independent priors a linear combination of $\hat{\lambda}_{1N}$ and $\hat{\lambda}_{1A}$ is better than $\hat{\lambda}_{1A}$, although not the best.

Proposition 6.1. Assume that λ_1 and λ_2 are independent a priori and the prior distribution of λ_i is Gamma (α,β), $i = 1,2$. Let $\hat{\lambda}_{1s} = a\hat{\lambda}_{1N} + (1-a)\hat{\lambda}_{1A}$, $0 \le a \le 1$. Then

(6.2) $$\min_a EE[(\hat{\lambda}_{1s} - \lambda_1)^2|\lambda_1,\lambda_2] = EE[(\hat{\lambda}^0_{1s} - \lambda_1)^2|\lambda_1,\lambda_2] ,$$

where

(i) $\hat{\lambda}^0_{1s} = a_0\hat{\lambda}_{1N} + (1-a_0)\hat{\lambda}_{1A}$ (assume that $\hat{\lambda}_{1N} \neq \hat{\lambda}_{1A}$)

$$a_0 = \begin{cases} 0, & \text{if } a_{00} \le 0 \\ a_{00}, & \text{if } 0 < a_{00} \le 1 \\ 1, & \text{if } a_{00} > 1 \end{cases}$$

$$a_{00} = [(n_1\bar{Y}_1 + \alpha)/(\beta + n_1) - \hat{\lambda}_{1A}]/(\hat{\lambda}_{1N} - \hat{\lambda}_{1A})$$

(ii) $\hat{\lambda}^0_{1s} = \hat{\lambda}_{1N} = \hat{\lambda}_{1A}$ (if $\hat{\lambda}_{1N} = \hat{\lambda}_{1A}$) .

The next proposition shows that the Bayes estimator for λ_1 does not depend on $\bar{Y}_2$ under the assumption of independent priors.

Proposition 6.2. Assume that λ_1 and λ_2 are independent a priori and the prior distribution of λ_i is Gamma (α,β), $i = 1,2$. Let $\hat{\lambda}_1 = \hat{\lambda}_1(\bar{Y}_1,\bar{Y}_2)$, a function of $\bar{Y}_1$ and $\bar{Y}_2$. Then

$$(6.3) \quad \min_{\hat{\lambda}_1} EE[(\hat{\lambda}_1 - \lambda_1)^2|\lambda_1,\lambda_2] = EE[(\hat{\lambda}_{10} - \lambda_1)^2|\lambda_1,\lambda_2] ,$$

where

$$\hat{\lambda}_{10} = (n_1\bar{Y}_1 + \alpha)/(n_1 + \beta) .$$

Following Propositions 6.1 and 6.2, we have the following result.

Proposition 6.3. Assume that λ_1 and λ_2 are independent a priori and the prior distribution of λ_i is Gamma (α,β), $i = 1,2$. Then

$$(6.4) \quad EE[(\hat{\lambda}_{10} - \lambda_1)^2|\lambda_1,\lambda_2] < EE[(\hat{\lambda}^0_{1s} - \lambda_1)^2|\lambda_1,\lambda_2] < EE[(\hat{\lambda}_{1N} - \lambda_1)^2|\lambda_1,\lambda_2],$$

where $\hat{\lambda}_{10}$, $\hat{\lambda}^0_{1s}$ and $\hat{\lambda}_{1N}$ are indicated respectively in (6.3), (6.2) and (6.1).

APPENDIX A

Appendix A gives proofs for some theorems in Chapter 1.

(i) Proof of Theorem 2.1: Before we proceed to prove Theorem 2.1, we need the following lemmas.

Lemma A.1.

$$\binom{k}{k} + \binom{k+1}{k} + \ldots + \binom{n}{k} = \binom{n+1}{k+1} ,$$

where $n \geq k$. Both n and k are non-negative integers.

Proof: Use induction on n.

Lemma A.2.

$$\sum_{k=0}^{n} k(k+1)(k+2)\ldots(k+m) = (m+1)! \binom{n+m+1}{m+2} .$$

Proof: Divide both sides by $(m+1)!$ and use Lemma A.1.

Lemma A.3.

$$\sum_{k=0}^{n} k^2(k+1)(k+2)\ldots(k+m) = (m+1)![n\binom{n+m+1}{m+2} - \binom{n+m+1}{m+3}] .$$

Proof: The left hand side $= (m+1)! \sum_{k=0}^{n} k\binom{k+m}{m+1}$

$= (m+1)![n\binom{n+m+1}{m+2} - \sum_{k=2}^{n}\binom{k+m}{m+2}] = (m+1)![n\binom{n+m+1}{m+2} - \binom{n+m+1}{m+3}]$

$=$ the right hand side. Q.E.D.

Lemma A.4. Consider the equation $n_1 + n_2 + \ldots + n_t = n$, $t \geq 2$, n is a fixed positive integer and n_i's are integer variables ≥ 0. Then there are $\binom{n+t-1}{t-1}$ possible solutions.

Proof: See Feller (1968), p. 38.

Lemma A.5. Consider the equation $n_1+n_2+\ldots+n_t = n$, $t \geq 2$, n is a fixed positive integer and n_i's are integer variables ≥ 0. Then among all possible solutions, each n_i takes

$$\binom{n+t-2}{t-2} \text{ 0's}$$

$$\binom{n+t-3}{t-2} \text{ 1's}$$

$$\vdots$$

$$\binom{n-k+t-2}{t-2} \text{ k's}$$

$$\vdots$$

$$\binom{t-2}{t-2} \text{ n .}$$

Proof: Consider $\{n_e, e = 1,2,\ldots,t | \sum_{e=1}^{t} n_e = n\} = \bigcup_{k=0}^{n} A_k$, where $A_k = \{n_e, e = 1,2,\ldots,t | \sum_{e=1}^{t-1} n_e = n-k, n_t = k\}$, $A_k \cap A_k' = \phi$ for $k \neq k'$. n_t takes k from A_k, and by Lemma A.4 there are $\binom{n-k+t-2}{t-2}$ ways that A_k can happen. Hence n_t takes $\binom{n-k+t-2}{t-2}$ k's among all possible solutions. Because of symmetric property, the result holds for other n_i's. Q.E.D.

Now, we are ready to prove Theorem 2.1. To prove (i), consider the equation $\sum_{i=1}^{t} n_i = n$. It follows that $\sum_{i=1}^{t} E(n_i) = n$. Because of symmetric property, $En_i = En_j$ for all $i \neq j$. Hence $E(n_i) = n/t$ for all i. To prove (ii), consider the following equation which follows from Lemmas A.4 and A.5.

$$\text{(A.1)} \qquad P(n_i = k) = \binom{n-k+t-2}{t-2} / \binom{n+t-1}{t-1}, \quad k = 0,1,2,\ldots,n .$$

Following from (A.1), we have

$$(A.2) \qquad E\, n_i^2 = \sum_{k=0}^{n} k^2 \binom{n-k+t-2}{t-2} / \binom{n+t-1}{t-1} .$$

Using Lemma A.1 it can be shown that

$$\sum_{k=0}^{n} k^2 \binom{n-k+t-2}{t-2} = \sum_{k=0}^{n} \sum_{e=0}^{n-k} k^2 \binom{e+t-3}{t-3}$$

$$(A.3) \qquad = \sum_{j=0}^{n} \sum_{e=0}^{j} (n-j)^2 \binom{e+t-3}{t-3} .$$

Applying Lemmas A.2 and A.3, it can be shown that

$$\sum_{j=0}^{n} \sum_{e=0}^{j} (n-j)^2 \binom{e+t-3}{t-3}$$

$$(A.4) \qquad = n^2 \binom{n+t-1}{t-1} - n(t-1)\binom{n+t-1}{t} - (t-1)\binom{n+t-1}{t+1} .$$

Following from (A.2), (A.3) and (A.4), we have

$$(A.5) \qquad E(n_i^2) = n^2/t - (t-1)n(n-1)/[t(t+1)] .$$

$\mathrm{Var}(n_i)$ stated in Theorem 2.1 follows from (A.5). To prove (iii), consider that $\mathrm{Var}(\sum n_i) = 0$. It follows that $\sum_i \mathrm{Var}(n_i) + 2 \sum_{i<j} \mathrm{Cov}(n_i, n_j) = 0$. From symmetry, it follows that $t\, \mathrm{Var}(n_i) + t(t-1)\mathrm{Cov}(n_i, n_j) = 0$. The result follows immediately. We finish the proof of Theorem 2.1.

(ii) Proof of Theorem 2.2.: Before we proceed to prove Theorem 2.2, we need the following lemma.

Lemma A.6. Assume that $(q_1, q_2, \ldots, q_t)$ is uniform on the simplex $(q_i > 0, \ \sum q_i = 1)$. Then

(i) $E(q_i) = 1/t$, for all i,

(ii) $\mathrm{Var}(q_i) = (t-1)/[t^2(t+1)]$, for all i,

(iii) $\mathrm{Cov}(q_i, q_j) = -1/[t^2(t+1)]$, for all $i \neq j$.

Proof: The uniform distribution of $(q_i, i = 1,2,\ldots,t)$ on the simplex is a special case of the $t-1$ variate Dirichlet distribution (when its parameters are all equal to 1). (See Wilks (1962), p.177-79). The result follows. Q.E.D.

Now, we are ready to prove Theorem 2.2. We note that $E(n_i|q_i) = nq_i$, $Var(n_i|q_i) = nq_i(1-q_i)$ and $Cov(n_i,n_j|q_i,q_j) = -nq_iq_j$. Applying Lemma A.6 and the following formulas:

$$
\begin{aligned}
&E(n_j) = EE(n_i|q_i) \ , \\
&Var(n_i) = Var\ E(n_i|q_i) + E\ Var(n_i|q_i) \ , \\
&E(n_in_j) = E_{q_i} E_{q_j} Cov(n_i,n_j|q_i,q_j) \\
&\qquad + E_{q_i} E_{q_j} [E(n_i|q_i,q_j) \cdot E(n_j|q_i,q_j)] \ ,
\end{aligned}
\tag{A.6}
$$

we can get $E(n_i)$, $Var(n_i)$ and $Cov(n_i,n_j)$ as stated in the theorem. Since the 1st and the 2nd moments of n_i's are one-to-one correspondence with $E(n_i)$, $Var(n_i)$ and $Cov(n_i,n_j)$, (ii) of Theorem 2.2 follows immediately. We finish the proof of Theorem 2.2.

(iii) Proof of Theorem 2.3: We note that $E(X|n_i) = \Sigma\, n_ip_i$ and $Var(X|n_i) = \Sigma\, n_ip_i(1-p_i)$. Applying Theorem 2.1 and (A.6), we get the result Q.E.D.

(iv) Proof of Corollary 2.1: We note that $E(X|q_i)$ and $Var(X|q_i)$ are the same as EX and Var X stated in Theorem 2.4. Applying Lemma A.6 and (A.6), we get the result. Q.E.D.

APPENDIX B

Appendix B gives proofs for some lemmas in Chapter 3.

(i) Proof of Lemma 2.2: Consider that $\partial D(p,\tau)/\partial p = 2ap + b\tau + d$, $\partial D(p,\tau)/\partial \tau = bp + 2c\tau + e$, $\partial^2 D(p,\tau)/\partial p^2 = 2a$, $\partial^2 D(p,\tau)/\partial \tau^2 = 2c$ and $\partial^2 D(p,\tau)/\partial p \partial \tau = b$. It follows that

$$\det \begin{bmatrix} \partial^2 D(p,\tau)/\partial p^2 & \partial^2 D(p,\tau)/\partial p \partial \tau \\ \partial^2 D(p,\tau)/\partial p \partial \tau & \partial^2 D(p,\tau)/\partial \tau^2 \end{bmatrix} = 4ac - b^2 .$$

When $n_2 \geq 2$, $4ac - b^2 > 0$. Setting $\partial D(p,\tau)/\partial p = \partial D(p,\tau)/\partial \tau = 0$ and solving for p and τ, we get $p = \frac{1}{2}$ and $\tau = 0$. It is clear that $(\frac{1}{2}, 0) \in R$. Q.E.D.

(ii) Proof of Lemma 2.3: When $p = 0$, $D(p,\tau) = c\tau^2 + e\tau$. The result follows immediately Q.E.D.

(iii) Proof of Lemma 2.4: If $\tau = 1-p$, then $D(p,\tau) = a'p^2+b'p+c'$, where $a' = (n_2^2+2n_1n_2+n_1n_2^2)/[n_1(n_1+ n_2)^2]$, $b' = -(n_2^2+2n_1n_2+2n_1n_2^2)/[n_1(n_1+n_2)^2]$ and $c' = n_2^2/(n_1+n_2)^2$. It follows that $(-b'+ \sqrt{b'^2-4a'c'})/(2a') = 1$ and $(-b'- \sqrt{b'^2- 4a'c'})/(2a') = p^*$. The result follows immediately. Q.E.D.

(iv) Proof of Lemma 2.5: If $\tau = -p$, then $D(p,\tau) = sp^2 - tp$, where $s = (n_2^2+ 2n_1n_2+ n_1n_2^2)/[n_1(n_1+ n_2)^2]$ and $t = (n_2^2+ 2n_1n_2)/[n_1(n_1+n_2)^2]$. The result follows immediately. Q.E.D.

(v) Proof of Lemma 2.6: If $p = 1$, then $D(p,\tau) = n_2(n_2-1)\tau^2/(n_1+ n_2)^2 - n_2\tau/(n_1+ n_2)^2$. The result follows immediately. Q.E.D.

APPENDIX C

Appendix C gives proofs for some lemmas in Chapter 4.

(i) Proof of Lemma 2.1: From Table 2.1 and (2.6), it is easy to see that

$$R(\tilde{p}_1) = aP(\tilde{p}_1 > \delta \text{ and } Y_1 \text{ being } \alpha) + bP(\tilde{p}_1 \leq \delta \text{ and } Y_1 \text{ being } \alpha) + cP(\tilde{p}_1 > \delta \text{ and } Y_1 \text{ being } \beta) + dP(\tilde{p}_1 \leq \delta \text{ and } Y_1 \text{ being } \beta) .$$

Note that $\tilde{p}_1$ and Y_1 are independent and $P(Y_1 \text{ being } \alpha) = p_1$. After some straightforward algebra, we have the result. Q.E.D.

(ii) Proof of Lemma 4.1: (i) First of all, we show that if $c_{1j} = c_{1k}$, $j \neq k$, then the $2 \times s$ case will degenerate into the $2 \times s-1$ case. Suppose $c_{1j} = c_{1k}$, $j \neq k$. Consider that

(C.1) Expected loss of $j = p_1 c_{1j} + q_1 c_{2j}$

(C.2) Expected loss of $k = p_1 c_{1k} + q_1 c_{2k}$.

Then, (C.1) $\leq$ (C.2) for all p_1 if $c_{2j} \leq c_{2k}$; (C.1) $>$ (C.2) for all p_1 if $c_{2j} > c_{2k}$. Either j is dominated by k, or vice versa. (ii) Similarly, if $c_{2j} = c_{2k}$, $j \neq k$, then the $2 \times s$ case will degenerate into the $2 \times s-1$ case. (iii) Suppose that c_{1j}'s are all unequal and c_{2j}'s are all unequal.

We put c_{ij}'s in order of magnitude from the smallest to the largest as in Condition (i) in (4.1). We are going to show if Condition (ii) in (4.1) does not hold, then the $2 \times s$ case will degenerate into the $2 \times \gamma$ case, $\gamma \leq s-1$. Suppose that $c_{2i_s} \geq c_{2i_{s-1}}$. Consider that

(C.3) Expected loss of $i_{s-1} = p_1 c_{1i_{s-1}} + q_1 c_{2i_{s-1}}$

(C.4) Expected loss of $i_s = p_1 c_{1i_s} + q_1 c_{2i_s}$.

It follows that (C.3) < (C.4); i.e., i_s is dominated by i_{s-1}. Similar results hold for other cases. Q.E.D.

(iii) Proof of Lemma 4.2: For $j = 2,\ldots,s-1$, $p_1 c_{1j} + q_1 c_{2j} \leq p_1 c_{1e} + q_1 c_{2e}$, $e = 1,2,\ldots,s$, $e \neq j \Leftrightarrow c_{2j} - c_{2e} \leq p_1(c_{1e} - c_{1j} + c_{2j} - c_{2e})$ $e = 1,2,\ldots,s$, $e \neq j \Leftrightarrow$

(C.5) $0 < c_{2j} - c_{2e} \leq p_1(c_{1e} - c_{1j} + c_{2j} - c_{2e})$, $e \geq j+1$, $j = 2,\ldots,s-1$

and

(C.6) $c_{2e} - c_{2j} \geq p_1(c_{1j} - c_{1e} + c_{2e} - c_{2j}) > 0$, $e \leq j-1$, $j = 2,\ldots,s-1$.

We claim that (C.5) $\Leftrightarrow$

(C.7) $0 < c_{2j} - c_{2j+1} \leq p_1(c_{1j+1} - c_{1j} + c_{2j} - c_{2j+1})$, $j = 2,\ldots,s-1$.

It is clear that (C.5) $\Rightarrow$ (C.7). To show that (C.7) $\Rightarrow$ (C.5), consider that (C.7) implies the following:

(C.8) $0 < c_{2j+1} - c_{2j+2} \leq p_1(c_{1j+2} - c_{1j+1} + c_{2j+1} - c_{2j+2})$.

From (C.7) and (C.8), it follows that

$$c_{2j} - c_{2j+2} \leq p_1(c_{1j+2} - c_{1j} + c_{2j} - c_{2j+2}) .$$

If we do this in appropriate steps, we can show that (C.5) holds. Similarly, (C.6) $\Leftrightarrow$

(C.9) $c_{2j-1} - c_{2j} \geq p_1(c_{1j} - c_{1j-1} + c_{2j-1} - c_{2j})$, $j = 2,\ldots,s-1$.

(i) in (4.4) follows from (C.7) and (C.9). Similar arguments for (ii) and (iii) in (4.4). Q.E.D.

(iv) Proof of Lemma 4.3: It is easy to see that

$$R(\tilde{p}_1) = c_{11}P(\tilde{p}_1 > \delta_1)p_1 + c_{21}P(\tilde{p}_1 > \delta_1)(1-p_1) + \sum_{j=2}^{s-1} c_{1j}P(\delta_j < \tilde{p}_1 \leq \delta_{j-1})p_1$$
$$+ \sum_{j=2}^{s-1} c_{2j}P(\delta_j < \tilde{p}_1 \leq \delta_{j-1})(1-p_1) + c_{1s}P(\tilde{p}_1 \leq \delta_{s-1})p_1$$
$$+ c_{2s}P(\tilde{p}_1 \leq \delta_{s-1})(1-p_1) .$$

After some messy but straightforward algebra, we have the result. Q.E.D.

(v) Proof of Lemma 5.1: It is easy to see that

$$R(\tilde{p}_{1j}) = \sum_{j=1}^{r} c_{j1}p_{1j}P(\sum_{j=1}^{r-1} \lambda_j \tilde{p}_{1j} > \lambda) + \sum_{j=1}^{r} c_{j2}p_{1j}P(\sum_{j=1}^{r-1} \lambda_j \tilde{p}_{1j} \leq \lambda).$$

After some straightforward algebra, we have the result. Q.E.D.

(vi) Proof of Lemma 6.1: It is easy to see that $R(\tilde{\mu}_1) = aP(\tilde{\mu} \leq c \text{ and } Y_1 > c) + bP(\tilde{\mu}_1 > c \text{ and } Y_1 \leq c)$. Note that $\tilde{\mu}_1$ and Y_1 are independent. The result follows. Q.E.D.

APPENDIX D

Appendix D gives proofs for propositions in Chapter 7.

(i) Proof of Proposition 2.1: It is easy to see that given (μ_1,μ_2),

$$E(\hat{\mu}_{1N}- \mu_1)^2 = \sigma^2/n_1 ,$$

$$E(\hat{\mu}_{1A}- \mu_1)^2 = \sigma^2/(n_1+ n_2) + [n_2^2/(n_1+ n_2)^2](\mu_2- \mu_1)^2 .$$

Also, it is easy to see that

$$E(\mu_2- \mu_1)^2 = (\mu_1^0- \mu_2^0)^2 + (\sigma_1^{0^2} + \sigma_2^{0^2} - 2\sigma_{12}^0) .$$

Therefore,

$$EE[(\hat{\mu}_{1A}- \mu_1)^2|\mu_1,\mu_2] = \sigma^2/(n_1+n_2)+[n_2^2/(n_1+ n_2)^2][(\mu_1^0-\mu_2^0)^2 + (\sigma_1^{0^2} + \sigma_2^{0^2} - 2\sigma_{12}^0)], \tag{D.1}$$

$$EE[(\hat{\mu}_{1N}- \mu_1)^2|\mu_1,\mu_2] = \sigma^2/n_1 .$$

(2.2) follows from (D.1). Q.E.D.

(ii) Proof of Proposition 2.2: Before we proceed to prove Proposition 2.2, we need the following lemma.

Lemma D.1. Suppose that given $\underset{\sim}{\mu}$, $\underset{\sim}{Y} \sim N(\underset{\sim}{\mu}, \mathfrak{L})$, where $\mathfrak{L}$ is known and positive-definite. Suppose that the prior distribution of $\underset{\sim}{\mu}$ is $N(\underset{\sim}{\mu}^0, \mathfrak{L}^0)$, where $\mathfrak{L}^0$ is positive-definite. Then the posterior distribution of $\underset{\sim}{\mu}$ given $\underset{\sim}{Y}$ is $N[\underset{\sim}{\mu}^*, (\mathfrak{L}^{0^{-1}} + \mathfrak{L}^{-1})^{-1}]$, where $\underset{\sim}{\mu}^* = (\mathfrak{L}^{0^{-1}} + \mathfrak{L}^{-1})^{-1}(\mathfrak{L}^{0^{-1}} + \mathfrak{L}^{-1})^{-1}(\mathfrak{L}^{0^{-1}} \underset{\sim}{\mu}^0 + \mathfrak{L}^{-1} \underset{\sim}{Y})$.

Proof: See DeGroot (1970) p. 175-76.

Now, we are ready to prove Proposition 2.2. It is easy to see

that $\hat{\mu}_1^*$ in (2.3) is $E(\mu_1|\bar{Y}_1, \bar{Y}_2)$. We note that

$$\begin{bmatrix}\bar{Y}_1\\ \bar{Y}_2\end{bmatrix} \sim N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma^2/n_1 & 0\\ 0 & \sigma^2/n_2\end{bmatrix}\right).$$

Then we can apply Lemma D.1 to compute $E(\mu_1|\bar{Y}_1,\bar{Y}_2)$. After messy but straightforward algebra, we get the expression for μ_1^* as indicated in (2.4). Q.E.D.

(iii) Proof of Proposition 3.1: For given $\underset{\sim}{B}$, we have shown (see Lemma 3.5 in Chapter 2) that

$$\begin{aligned} E(\hat{Y}_0-Y_0)^2 &= \sigma^2 + \sigma^2\, \underset{\sim}{X}_{10}'(X_1'X_1)^{-1}\, \underset{\sim}{X}_{10} + \sigma^2\, \underset{\sim}{Z}_0'\, W^{-1}\underset{\sim}{Z}_0 ,\\ E(\hat{\hat{Y}}_0-Y_0)^2 &= \sigma^2 + \sigma^2\, \underset{\sim}{X}_{10}'(X_1'X_1)^{-1}\, \underset{\sim}{X}_{10} + (\underset{\sim}{Z}_0'\, \underset{\sim}{B}^2)^2 . \end{aligned} \tag{D.2}$$

It is easy to see that

$$E(\underset{\sim}{Z}_0'\, \underset{\sim}{B}^2)^2 = \underset{\sim}{Z}_0'\, \Sigma_{22}^0\, \underset{\sim}{Z}_0 + \underset{\sim}{Z}_0'\, \underset{\sim}{B}_2^0\, \underset{\sim}{B}_2^{0\prime}\underset{\sim}{Z}_0 . \tag{D.3}$$

(3.5) follows from (D.2) and (D.3). Q.E.D.

(iv) Proof of Proposition 3.2: It is easy to see that $E(\underset{\sim}{B}|\underset{\sim}{Y})$ minimizes $E(\tilde{\underset{\sim}{B}} - \underset{\sim}{B})(\tilde{\underset{\sim}{B}} - \underset{\sim}{B})'$ in the positive semi-definite sense. Following from (3.6), it is easy to see that $\begin{bmatrix}\underset{\sim}{B}\\ \underset{\sim}{Y}\end{bmatrix}$ has a multivariate normal distribution with

$$E\begin{bmatrix}\underset{\sim}{B}\\ \underset{\sim}{Y}\end{bmatrix} = \begin{bmatrix}\underset{\sim}{B}^0\\ X\underset{\sim}{B}^0\end{bmatrix} \text{ and } \operatorname{var}\begin{bmatrix}\underset{\sim}{B}\\ \underset{\sim}{Y}\end{bmatrix} = \begin{bmatrix}\Sigma^0 & \Sigma^0 X'\\ X\Sigma^0 & X\Sigma^0 X' + \Sigma\end{bmatrix} .$$

(3.7) follows immediately. Q.E.D.

(v) Proof of Proposition 3.3: Consider

$$E(Y_0 - \tilde{Y}_0)^2 = E(\underset{\sim}{X}_0' \underset{\sim}{B} + e_0 - \underset{\sim}{X}_0' \tilde{\underset{\sim}{B}})^2 = \sigma^2 + E(\underset{\sim}{X}_0' \underset{\sim}{B} - \underset{\sim}{X}_0' \tilde{\underset{\sim}{B}})^2$$

$$= \sigma^2 + \underset{\sim}{X}_0' E(\underset{\sim}{B} - \tilde{\underset{\sim}{B}})(\underset{\sim}{B} - \tilde{\underset{\sim}{B}})' \underset{\sim}{X}_0 .$$

By Proposition 3.2, $\underset{\sim}{X}_0' E(\underset{\sim}{B} - \tilde{\underset{\sim}{B}})(\underset{\sim}{B} - \tilde{\underset{\sim}{B}})' \underset{\sim}{X}_0$ is minimized when $\tilde{\underset{\sim}{B}} = \hat{\underset{\sim}{B}}^* = E(\underset{\sim}{B}|\underset{\sim}{Y})$. The result follows. Q.E.D.

(vi) Proof of Proposition 4.1: We want to apply (3.5) to this particular model (4.2). It is easy to see that in this particular model,

$$\underset{\sim}{Z}_0 = \underset{\sim}{X}_{20} - X_2'X_1(X_1'X_1)^{-1}\underset{\sim}{X}_{10} = n_2/(n_1 + n_2) ,$$

$$W = (X_2'X_2) - X_2'X_1(X_1'X_1)^{-1}X_1'X_2 = n_1n_2/(n_1 + n_2) ,$$

$$\Sigma_{22}^0 = \sigma_1^{0^2} + \sigma_2^{0^2} - 2\sigma_{12}^0 ,$$

$$B_2^0 = \mu_1^0 - \mu_2^0 .$$

The result follows. Q.E.D.

(vii) Proof of Proposition 4.2: We want to apply (3.7) to this particular model (4.11) to compute $\hat{\gamma}^*$. In this particular model, it is easy to see that

$$X \Sigma^0 X' = \begin{bmatrix} \sigma_1^{0^2} & \sigma_{12}^0 \\ \sigma_{12}^0 & \sigma_2^{0^2} \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \sigma^2/n_1 & 0 \\ 0 & \sigma^2/n_2 \end{bmatrix} .$$

It follows that

$$(X \Sigma^0 X' + \Sigma)^{-1} = \frac{1}{k} \begin{bmatrix} \sigma_2^{0^2} + \sigma^2/n_2 & -\sigma_{12}^0 \\ -\sigma_{12}^0 & \sigma_1^{0^2} + \sigma^2/n_1 \end{bmatrix},$$

where

$$k_1 = (\sigma_1^{0^2} + \sigma^2/n_1)(\sigma_2^{0^2} + \sigma^2/n_2) - (\sigma_{12}^0)^2 .$$

Also, it is easy to see that in this particular model,

$$\Sigma^{0} X' = \begin{bmatrix} \sigma_1^{0^2} & \sigma_{12}^{0} \\ \sigma_{12}^{0}-\sigma_1^{0^2} & \sigma_2^{0^2} - \sigma_{12}^{0} \end{bmatrix} \quad \text{and} \quad \underset{\sim}{Y} - X\underset{\sim}{B}^{0} = \begin{bmatrix} \bar{Y}_1 - \mu_1^{0} \\ \bar{Y}_2 - \mu_2^{0} \end{bmatrix}.$$

After some messy but straightforward algebra, we can find $\hat{Y}^{*} = E[Y|\bar{Y}_1, \bar{Y}_2]$ as stated in (4.14). Q.E.D.

(viii) Proof of Proposition 5.1: The p.d.f. of (X_1,X_2) given (p_1,p_2) is

$$f(x_1,x_2|p_1,p_2) = \binom{n_1}{x_1} p_1^{x_1}(1-p_1)^{n_1-x_1}\binom{n_2}{x_2} p_2^{x_2}(1-p_2)^{n_2-x_2},$$

(D.4)
$$x_i = 0,1,\ldots,n_i, \quad i = 1,2.$$

The p.d.f. of (p_1,p_2) is

$$g(p_1,p_2) = \{\Gamma(\nu_1+\nu_2)/[\Gamma(\nu_1)\Gamma(\nu_2)]\}^2 p_1^{\nu_1-1}(1-p_1)^{\nu_2-1} p_2^{\nu_1-1}(1-p_2)^{\nu_2-1},$$

(D.5)
$$0 < p_i < 1, \quad i = 1,2.$$

It follows that the posterior p.d.f. of (p_1,p_2) given (X_1,X_2) is

$$k(p_1,p_2|X_1,X_2) = \prod_{i=1}^{2}\{\Gamma(n_i+\nu_1+\nu_2)/[\Gamma(x_i+\nu_1)\Gamma(n_i+\nu_2-x_i)]\} \cdot p_i^{x_i+\nu_1-1}(1-p_i)^{n_i+\nu_2-x_i-1}.$$

It follows that

(D.6) $$E(p_1|X_1,X_2) = (X_1+\nu_1)/(n_1+\nu_1+\nu_2).$$

It is clear that $\min_a EE[(\hat{p}_{1s}-p_1)^2|p_1,p_2] = \min_a E[(\hat{p}_{1s}-p_1)^2|X_1,X_2]$.

Let $T(a) = E[(\hat{p}_{1s}- p_1)^2|X_1,X_2]$. Then

$$dT(a)/da = 2E[(a\hat{p}_{1N} + (1-a)\hat{p}_{1A} - p_1)(\hat{p}_{1N} - \hat{p}_{1A})|X_1,X_2] ,$$

$$d^2T(a)/da^2 = 2E[(\hat{p}_{1N} - \hat{p}_{1A})^2|X_1,X_2] > 0 \text{ (assume } \hat{p}_{1N} \neq \hat{p}_{1A}).$$

Setting $dT(a)/da = 0$ and solve for a, we get $a = a_{00}$ as stated in the proposition. Q.E.D.

(ix) Proof of Proposition 5.2: It is clear that $\min_{\hat{p}} EE[(\hat{p} - p_1)^2|p_1,p_2] = \min_{\hat{p}} E[(\hat{p} - p_1)^2|X_1,X_2]$. However, $E(p_1|X_1,X_2)$ minimizes $E[(\hat{p} - p_1)^2|X_1,X_2]$, where $E(p_1|X_1,X_2)$ is as indicated in (D.6). Q.E.D.

(x) Proof of Proposition 5.3: It follows directly from Propositions 5.1 and 5.2.

(xi) Proof of Proposition 6.1: Given (λ_1,λ_2), $Z_i = n\bar{Y}_i \sim$ Poisson (λ_i'), where $\lambda_i' = n_i\lambda_i$, $i = 1,2$. Also, $\lambda_i \sim$ Gamma$(\alpha,\beta) \Rightarrow \lambda_i' \sim$ Gamma(α,β_i), where $\beta_i = \beta/n_i$, $i = 1,2$. It is easy to see that λ_1' and λ_2' are independent. It follows that the posterior p.d.f. of (λ_1', λ_2') given (Z_1,Z_2) is

$$h(\lambda_1',\lambda_2'|z_1,z_2) = \pi_{i=1}^{2}[(\beta_i+1)^{z_i+\alpha}/\Gamma(z_i+\alpha)]\lambda_i'^{z_i+\alpha-1} e^{-(\beta_i+1)\lambda_i'} ,$$

(D.7) $$\lambda_i' > 0 .$$

It follows that

(D.8) $$E(\lambda_1|Z_1,Z_2) = E(\lambda_1'|Z_1,Z_2)/n_1 = (n_1\bar{Y}_1 + \alpha)/(n_1 + \beta) .$$

It is clear that $\min_a EE[(\hat{\lambda}_{1s} - \lambda_1)^2|\lambda_1,\lambda_2] = \min_a E[(\hat{\lambda}_{1s} - \lambda_1)^2|Z_1,Z_2]$. The rest of proof is similar to the proof of Proposition 5.1. Q.E.D.

(xii) Proof of Proposition 6.2: It is similar to the proof of Proposition 5.2.

(xiii) Proof of Proposition 6.3: It follows directly from Propositions 6.1 and 6.2.

REFERENCES

(1) Anderson, R. L., D. M. Allen and F. B. Cady. (1972). Selection of predictor variables in linear multiple regression. Statistical Papers in Honor of George W. Snedecor, The Iowa State University Press, 3 - 18.

(2) Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. Ann. Math. Statist. 15: 190 - 204.

(3) Bancroft, T. A. (1964). Analysis and inference for incompletely specified test(s) of significance. Biometrics. 20: 427 - 42.

(4) Bancroft, T. A. (1972). Some recent advances in inference procedures using preliminary tests of significance. Statistical Papers in Honor of George W. Snedecor, The Iowa State University Press, 19 - 30.

(5) Bozivich, H., T. A. Bancroft and H. O. Hartley. (1956). Power of analysis of variance test procedures for incompletely specified models. Ann. Math. Statist. 27: 1017 - 43.

(6) DeGroot, M. H. (1970). Optimal Statistical Decisions. McGraw-Hill, Inc.

(7) Feller, W. (1968). An Introduction to Probability Theory and Its Applications. Vol. 1 (3rd ed.). Wiley, New York.

(8) Han, Chien-Pai and T. A. Bancroft. (1968). On pooling means when variance is unknown. J. Amer. Statist. Assoc. 63: 1333 - 42.

(9) Huntsberger, D. V. (1955). A generalization of a preliminary testing procedure for pooling data. Ann. Math. Statist. 26: 734 - 743.

(10) Kale, B. K. and T. A. Bancroft. (1967). Inferences for some incompletely specified models involving normal approximations to discrete data. Biometrics. 23: 335 - 48.

(11) Mosteller, F. (1948). On pooling data. J. Amer. Statist. Assoc. 21: 231 - 42.

(12) Rao, C. R. (1965). Linear Statistical Inference and Its Applications. Wiley, New York.

(13) Schneider, R. (1970). A mean squared error criterion for selecting a subset of predictor variables. Ph.D. Thesis, University of Minnesota, Minneapolis.

(14) Toro-Vizcarrondo, C. E. and T. D. Wallace. (1968). A test of the mean square error criterion for restrictions in linear regression. J. Amer. Statist. Assoc. 63: 558 - 72.

(15) Wilks, S. S. (1962). Mathematical Statistics. Wiley, New York.